MITO: A Millimeter-Wave Dataset and Simulator for Non-Line-of-Sight Perception

Tara Boroushaki[*1], **Laura Dodds**[*1], **Cusuh Ham**[2], **Fadel Adib**[1,3]
tarab@mit.edu, ldodds@mit.edu, ham@adobe.com, fadel@mit.edu
[1] Massachusetts Institute of Technology, [2] Adobe Research, [3] Cartesian Syestems

Abstract

The ability to observe the world is fundamental to reasoning and making informed decisions on how to interact with the environment. However, optical perception can often be disrupted due to common occurrences like occlusions, which can pose challenges to existing vision systems. We present MITO, the first millimeter-wave (mmWave) dataset of diverse, everyday objects, collected using a UR5 robotic arm with two mmWave radars operating at different frequencies and an RGB-D camera. Unlike visible light, mmWave signals can penetrate common occlusions (e.g., cardboard boxes, fabric, plastic). MITO captures over 24 million mmWave frames and uses them to generate 550 high-resolution mmWave images in line-of-sight and non-light-of-sight (NLOS), as well as RGB-D images, segmentation masks, and raw mmWave signals, taken from 76 different objects. We develop an open-source simulation tool that can be used to generate synthetic mmWave images for any 3D triangle mesh. Finally, we establish benchmarks for NLOS segmentation and classification to demonstrate the utility of our dataset and simulator for enabling broader NLOS perception.

1 Introduction

Recent decades have witnessed many impressive advancements in computer vision, with state-of-the-art systems capable of high-accuracy object detection, segmentation, classification, and more. One of the main applications of these advances is in robotic manipulation and perception, where autonomous robots are expected to find and retrieve objects and navigate through complex environments spanning warehouses, manufacturing plants, and smart homes.

However, relying primarily on optical perception (e.g., cameras, LiDARs, etc) inherently limits existing systems to line-of-sight (LOS).[1] For example, a camera-based system cannot see inside a closed box to confirm e-commerce orders are correct or detect broken objects during shipping. Nor is it feasible for camera-based robots to efficiently plan complex, multi-step tasks in practical environments where required objects may be fully occluded.

Instead, it is possible to produce non-line-of-sight (NLOS) images using millimeter-wave (mmWave) radars [6]. Unlike visible light, mmWave wireless signals can traverse everyday occlusions such as cardboard, fabric, and plastic. This allows mmWave radars to image objects in non-line-of-sight, similar to how airport security scanners produce high-resolution NLOS images of passengers to detect hidden weapons. With the emergence of low-cost mmWave radars, it is possible to build pervasive autonomous systems with this NLOS imaging capability. For example, using mmWave radars attached to a robotic arm (Fig. 1), we can produce a NLOS mmWave image of a wrench within a closed box.

However, to achieve the many possible applications of NLOS imaging, we need to be able to *perceive* (e.g., segment, classify, etc) objects within mmWave images. Unfortunately, these images are fundamentally

[1] While some visible-light systems can rely on reflected laser light to image objects around the corner, they cannot image fully occluded items (e.g., within a closed box).

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

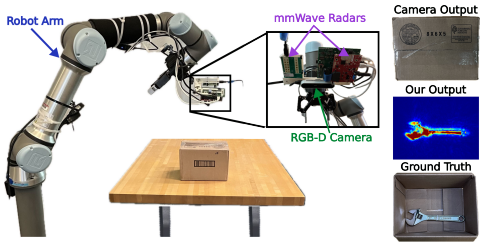


Figure 1: **MITO.** We use a robotic arm to move mmWave radars throughout the environment. While an RGB-D camera's output cannot see inside the box, we produce high-resolution non-line-of-sight mmWave images.

different from visible-light images in multiple ways: they 1) suffer from radar-specific artifacts (e.g., the much longer wavelength of mmWaves versus visible light–hundreds of nanometers–results in fundamentally different specular and diffuse properties of imaged objects), 2) have no color information, and 3) are complex-valued.[2] Therefore, since the majority of existing perception algorithms have not been trained on these mmWave images, they need to be adapted for NLOS perception. In some cases, we may require entirely new algorithms or models to enable mmWave-based perception to approach the accuracy of visual-light-based perception, as shown in this work.

Inspired by how early datasets (e.g., MSRC [22], Pascal [20], Berkeley 3D [28], and YCB [14]) paved the way for many advancements within computer vision, we believe that a mmWave image dataset of everyday objects is fundamental for enabling pervasive NLOS perception. To summarize, we make the following key contributions:

- **Dataset.** We create MITO, a novel dataset of mmWave images, using a custom-built robotic imaging system and signal processing pipeline. We capture over 24 million mmWave frames, and use them to generate 550 images of 76 everyday objects in both line-of-sight and non-line-of-sight. For each object, we capture mmWave images at multiple different frequencies. We also provide RGB-D images of the objects (when in LOS) and ground truth object segmentations.
- **Simulator.** We develop an open-source simulator that can produce synthetic mmWave images for any 3D triangle mesh. Since different materials interact differently with mmWave signals and, therefore, produce different images, we build two different modeling methods into our simulation, the combination of which can be used to simulate a wide range of material properties[3].
- **Non-line-of-sight experiments.** Using MITO, we establish baselines for two computer vision tasks in non-line-of-sight: 1)object segmentation, using the segment-anything model (SAM) [29] and a mmWave power-based prompter, and 2) shape classification, which is trained entirely on synthetic images and generalizes to real images.

2 Related Work

mmWave Datasets. Recognizing the importance of millimeter-wave perception, past research has made initial steps towards millimeter-wave datasets. However, the vast majority of them are focused on either imaging humans for action recognition [15, 16, 48, 39, 50], outdoor imaging for autonomous driving [11, 13, 23, 30, 36, 53, 37, 51, 47, 24], or imaging building interiors (walls, hallways) [12, 38]. For higher-resolution imaging, datasets are limited to domain-specific objects like guns and knives for TSA where detecting concealed weapons through body scanners is important for airport security applications [33, 41, 40]. Our work, however, is focused on generating a dataset of a larger, more diverse set of everyday objects for applications such as robotic manipulation and perception.

mmWave Image Simulators. Given the growth in millimeter-wave devices for networking and sensing, the vast majority of existing mmWave simulations have also focused on understanding channel characteristics (e.g., for 5G networks) [8, 35, 43, 26] or for imaging humans for action recognition and HCI applications [49, 54, 52]. However, [35, 43, 26] do not support small indoor objects, and [49, 54, 52] focus exclusively on human body modeling, rely on assumptions about scattering from limbs (which do not generalize to arbitrary objects). In principle, simulators such as [8, 40] could be used for generating mmWave images of everyday

[2] Interestingly, the values are dependent on not only the location of the object, but also the material and texture.

[3] The code and data will be added upon publication. A video demonstration is available here: https://www.youtube.com/watch?v=atAqefGl3Sc

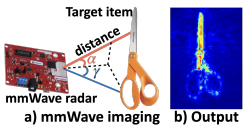


Figure 2: **mmWave Imaging** a) mmWave radars estimate range and angle-of-arrival to produce b) reflection maps.

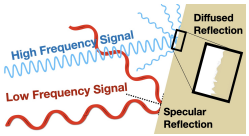


Figure 3: **Diffuse and Specular Reflections.** The frequency of a signal changes the type of reflections it experiences for the same surface.

Dataset	Targets	FCC Bandwidth	Publicly Available	Resolution
MITO	76 Everyday Objects	✓	✓	High
[33]	30 TSA Objects	✗	✗	High
[40]	9 Objects Categories	✗	✓	High
[15, 48, 16, 39, 50]	Humans	✓	✓	Low
[11, 13, 23, 30] [53, 37, 51, 24]	Vehicles & Driving Scenes	✓	✓	Low

Table 1: **mmWave Dataset Comparisons.**

objects, but they are extremely expensive, costing tens of thousands of dollars [8] and/or are specific to simulating metallic items, limiting their application to everyday objects with different properties [40].

In contrast, we develop an open-source simulator and demonstrated that it generates synthetic data that matches real-world collected data. We believe that this tool, combined with advances in vision algorithms, can lead to improvements in performance in non-line-of-sight scenarios.

mmWave Perception. Prior work has demonstrated the use of millimeter waves for various perception tasks, including detection and classification. However, these approaches are limited to specific domains, such as vehicles [31, 23, 44], pedestrians [25], weapons [41, 42], or entirely metallic objects [32]. Through a new dataset and simulator, our work aims to expand the application of mmWave perception to a more open domain consisting of diverse categories.

3 Background

mmWave Imaging. A millimeter-wave radar transmits wireless signals with mmWave wavelengths, captures their reflections, and uses them to generate 3D images (shown in Fig. 2a). It uses two core signal processing techniques. First, the radar estimates the distance of each reflection by comparing the frequency between the transmitted and received signals. Second, it measures the angle of reflection (in both azimuth and elevation) by comparing the received signals between different antennas. 3D images can then be formed by calculating the reflection power emanating from each voxel in 3D space.

As shown in Fig. 2b, the captured image can be represented as a heatmap, where red represents high intensity reflection and blue represents low intensity. It is important to note that this visualization is a 2D projection of the 3D image. Mathematically, since every sample from the radar is a complex number with a phase and an amplitude, we can combine them with the following equation:

$$I(x,y,z) = \sum_{k=1}^{K}\sum_{j=1}^{N} S_{j,k} e^{\frac{j2\pi d(x,y,z,k)}{\lambda_j}} \tag{1}$$

where $I(x,y,z)$ is the complex mmWave image at location (x,y,z), K is the number of antennas, N is the number of samples in each signal, $S_{j,k}$ is the j$^{\text{th}}$ sample of the received signal from the k$^{\text{th}}$ antenna, and λ_j is the wavelength of the signal at sample j, which is proportional to the speed of light c and the frequency f as $\lambda = \frac{c}{f}$. $d(x,y,z,k) = 2||(x,y,z) - p_k||$ is the round-trip distance from the k$^{\text{th}}$ antenna (at location p_k) to the point (x,y,z).

Resolution. One important characteristic of a mmWave image is its resolution. The resolution in range (depth) is determined by the *bandwidth*, or the range of frequencies that the signal covers. We compute the depth resolution $\delta_z = \frac{c}{2B}$ where c is the speed of light, and B is the bandwidth of the signal. The resolution in the horizontal and vertical dimensions is dependent on the *aperture*, or the distance between the first and last antenna in that dimension. $\delta_x = \frac{\lambda z_0}{2D_x}$ and $\delta_y = \frac{\lambda z_0}{2D_y}$ are the resolution in the x and y dimensions, respectively. z_0 is the range to the target, and D_x and D_y are the apertures in x and y dimensions, respectively.

Synthetic Aperture Radar (SAR) Imaging. As described above, it is desirable to increase the aperture to increase the image resolution. Instead of creating a large aperture through a large number of physical antennas on the radar, it is possible to perform SAR imaging. In this method, one set of transmitters/receivers are moved through the environment, taking measurements from different locations, and constructing a "synthetic aperture". These measurements can be combined through Eq. 1 to provide the same resolution as a physical aperture.

Impact of Frequency. The frequency used for imaging introduces an important tradeoff – higher frequencies result in higher resolution, but the signals experience more *attenuation*, or power loss, as they

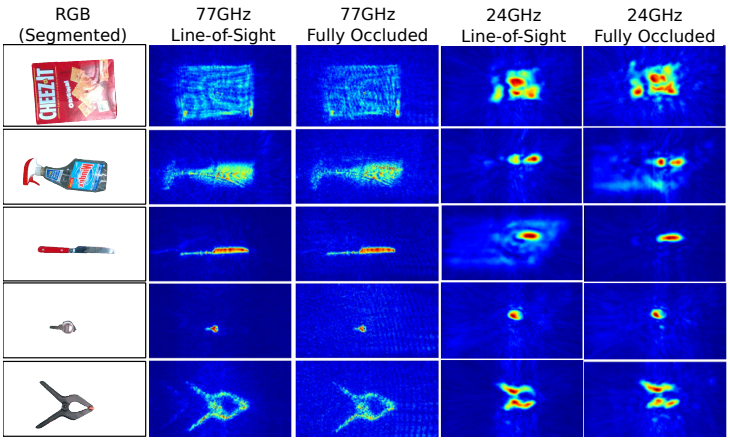


Figure 4: **Sample Images.** Example mmWave images from MITO.

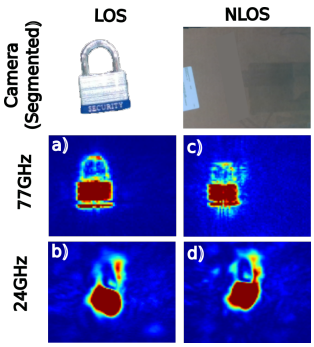


Figure 5: **Multi-Spectral Imaging.** a-b) show 77 GHz & 24 GHz images for a padlock in LOS. c-d) show the images in NLOS.

travel through materials [5]. In other words, high frequency signals (e.g., Terahertz) struggle with thick occlusions due to higher noise [17], but lower frequencies (e.g., WiFi) suffer from poor resolution [9]. Therefore, we choose multi-spectral (multi-frequency) mmWave signals for our dataset and simulator. Specifically, we select 24 GHz and 77 GHz since these are the bands approved by the FCC for industrial and commercial use [21], and most commercial mmWave radars operate within them.

Types of Reflections. mmWave signals primarily exhibit specular (mirror-like) reflections [34], illustrated in Fig. 3. Due to the significantly longer wavelengths of mmWaves compared to visible light (millimeters vs. nanometers), surfaces that appear diffuse (i.e., omnidirectional) for visible light (blue) often appear specular for mmWave signals (red). mmWaves also exhibit edge diffraction, or strong reflections when a wave strikes a sharp edge, scattering it in multiple directions. Edge diffraction can impact signal propagation and perception differently from diffuse or specular reflections. These reflection properties are one fundamental difference between mmWave signals and visible light that necessitates models and algorithms specifically designed for mmWave images.

4 MITO Dataset

We describe the process of creating MITO, which contains over 24 million mmWave frames where each frame is a complex-valued time series. These frames are used to generate 550 unique, high-resolution 3D mmWave SAR images of 76 diverse everyday objects. MITO also includes multi-spectral (77 GHz and 24 GHz) mmWave images, RGB-D images, and ground truth object segmentation masks. Table 1 provides an overview comparing MITO with other existing mmWave datasets.

Object Selection. MITO builds on the YCB dataset [14], a standard dataset of everyday objects used for robotic manipulation. We select 76 out of 80 objects from the YCB dataset (the remaining 4 were discarded due to not being able to obtain the exact object). Unlike previous mmWave datasets, these objects cover a diverse range of shapes, sizes, materials, and categories (additional details on object diversity can be found in the supplementary). The YCB dataset also provides 2D images and 3D object meshes, which we use to simulate mmWave images.

Data Collection Protocol. *Real-World mmWave Images.* Fig. 1 visualizes our data collection setup. We connect a 77 GHz radar (TI IWR1443Boost [6] and DCA1000EVM [1]), 24 GHz radar (Infineon Position2Go [4]), and Intel Realsense D415 [27] to an end-effector of a UR5e robotic arm [46] using custom-designed 3D-printed parts. Since the robot and radars are controlled from different computers, we synchronize them by setting both computers to the same NTP time server.

We place the objects in two setups that minimize background reflections in different ways. In the first, we fix each object onto a tripod by hot-gluing a 3D-printed nut to each object and screwing it onto a threaded tripod. Because there is no surface directly below the object, any strong reflections from the background are avoided. In the second, we place the object directly on a styrofoam block, since styrofoam is largely transparent to mmWave signals [7]. We also include experiments where the tripod/styrofoam is placed in the scene without any objects so that the impact of the tripod/styrofoam can be removed by subtracting the empty image from the main image [10].

At the start of an experiment, the robot moves the camera to the middle of the workspace to take an RGB-D image prior to any potential occlusion (i.e., in the NLOS scenario). Then, it moves the radars throughout

Simulator	Compatible Objects	Simulates Objects & Indoors	Material Diversity	Publicly Available	Cost
MITO	Everyday Objects	✓	✓	✓	Free
Sionna[26]	Cars, streets, buildings, etc	✗	✓	✓	Free
Ansys [8]	Everything	✓	✓	✗	10-50K$
[40]	Metalic objects	✓	✗	✗	Free
[49]	Humans	✗	✗	✗	–
[54]	Humans	✗	✗	✓	Free
[52]	Humans	✗	✗	✓	Free

Table 2: **mmWave Simulators Comparisons.**

the workspace, covering a rectangular space. The aperture of our experiments range from 55cm×22cm to 60cm×45cm, ensuring that each experiment uses an aperture which covers the full object. The radars continuously measure mmWave signals while the robot is moving, and the robot records its trajectory over time. After we complete an experiment in LOS, we run an experiment in NLOS without moving the object. We place a layer of cardboard over the object to simulate the case where the object was placed inside a box. We also collect experiments with four layers of fabric over the cardboard to simulate more dense occlusions.

Synthetic mmWave Images. In addition to the real-world data, we also provide a simulation tool to generate synthetic mmWave images, which can be used to increase the volume of data to aid in model training. We include at least 2 simulation images per object, along with the simulator for producing additional images.

Data Processing. To process a real-world mmWave image, we need to know the location where each measurement was taken. Thus, we interpolate the timestamped robot locations to match the timestamp of the radar measurements. The interpolated locations, in addition to the raw radar measurements, are then processed into a 3D mmWave image by applying Eq. 1. Example images can be seen in Fig. 4, which shows the segmented RGB images (1st column), 77 GHz LOS (2nd column), 77 GHz NLOS (3rd column), 24 GHz LOS (4th column), and 24 GHz NLOS (5th column) mmWave images. Additional examples, including an analysis of the impact of various occlusion types (such as plastic, drywall, and plywood), are provided in the supplementary materials.

We produce 550 3D, complex-valued mmWave images at two different frequencies (24 GHz & 77 GHz) in both LOS and NLOS settings using 24 million raw measurements and locations. We include the raw data in MITO to enable flexibility in leveraging and/or postprocessing the data using a custom pipeline. We also include images of objects at different angles and images with multiple objects.

For each mmWave image, we use its corresponding RGB-D image (in LOS) to provide a ground truth object segmentation mask. The segmentation masks are produced by using an interactive interface for the segment-anything model (SAM) [29] to select the points on the RGB picture as a prompt for the model. The human annotator continues selecting points until the mask covers the full object. We then use the camera's location to transform and align the RGB-D ground-truth mask with the mmWave image. In addition, we align the 3D ground-truth mesh model of each object with the corresponding depth images. This alignment enables richer annotations beyond segmentation, including 3D bounding boxes and full 6-DoF object pose labels. More details on the alignment process are in the supplementary.

Analysis of Different Frequencies. Recall from Sec. 3 that the frequency of the mmWave signal impacts both the resolution of the resulting image and the ability for the signal to traverse through occlusions. We analyze this phenomenon in real-world images in MITO in Fig. 5. Fig.5a-b show a 77 GHz and 24 GHz image of a padlock in LOS, respectively. The 77 GHz image is higher resolution, as expected. On the other hand, Fig. 5c-d show 77 GHz and 24 GHz images of the same object in NLOS (placed underneath cardboard). In this case, the 24 GHz image is very similar to the LOS case, while the 77 GHz shows some artifacts that were not present in LOS. This is because the 24 GHz signal travels through the cardboard relatively unaffected, while the 77 GHz is partially reflected and absorbed. We hope that by including two different frequencies, future algorithms can leverage both images to benefit from high-resolution and robustness to occlusions.

5 MITO Simulator

We describe our MITO mmWave image simulator. Unlike other mmWave simulators, MITO is compatible with diverse objects, accommodates variable material properties, and is open-source. We provide a comparison to other mmWave simulators in Table 2.

5.1 mmWave Simulation

Our simulator takes as input a 3D triangle mesh and radar locations to simulate mmWave signals from. The mesh files can be sourced from object datasets (e.g., YCB [14]), online repositories (e.g., 3D Warehouse, Turbo Squid), or custom-generated on smartphones (e.g., PolyCam iPhone app).

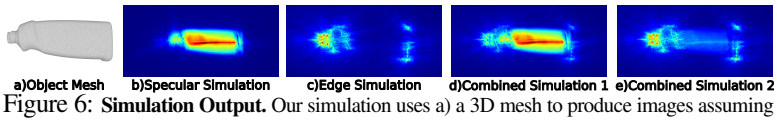


Figure 6: **Simulation Output.** Our simulation uses a) a 3D mesh to produce images assuming different reflection types: b) specular & c) edge. d/e) We combine these with different weights to simulate different materials.

Sim	25th	50th	75th
Comb	**85%**	**94%**	**98%**
Spec	68%	90%	97%
Edge	49%	71%	88%

Table 3: **Simulation 3D F-Score** for combined (Comb), specular (Spec), & edge simulations.

At a high level, our simulator works by estimating the raw signals expected at each of these locations, given reflections off the target object. Then, synthetic images are produced by feeding these simulated signals into the same image processing pipeline used for real world data (Eq. 1).

The process of estimating raw signals at a given location is as follows. (1) We find which mesh vertices in the mesh are visible from that location [3]. (2) For a given visible vertex, we measure the distance to the measurement location. Given the distance, we simulate the reflection from this vertex as:

$$s_j(l,v) = e^{-j\frac{4\pi|l-v|}{\lambda_j}} \tag{2}$$

where $s_j(l,v)$ is the j$^{\text{th}}$ sample of the reflection from location l to vertex v and back. (3) We sum the simulated reflections from all visible vertices V to produce the raw signal for this measurement location: $S_j = \sum_{v \in V} s_j(l,v)$ where S_j is the j$^{\text{th}}$ sample of the simulated signal from location l.

After repeating this for each measurement location, we feed the raw signals through our imaging processing pipeline to produce the final synthetic image. Note that we can rotate the mesh files to produce synthetic images at various angles. Additionally, recall from Sec. 3 that the object's material and texture may result in different types of reflections. To accurately model different objects, we introduce two variations of the simulation to account for different types of reflections.

Specular Reflections. When an object exhibits purely specular reflections, the signal will reflect off each vertex in only one direction. In some cases, the signal is reflected away from the radar's receivers, and therefore should not be added to S_j. Specifically, the signal will only be reflected back towards the radar if the surface normal is pointing towards the radar's location. Otherwise, the signal will be reflected away.

We update our above simulation to account for this by removing reflections where the angle between the normal ($\boldsymbol{n}$) and the vector from the radar to the vertex ($l-v$) is less than a threshold τ. Formally:

$$s_j(l,v) = \begin{cases} 0 & \cos^{-1}\left(\frac{\boldsymbol{n}\cdot(l-v)}{|\boldsymbol{n}||(l-v)|}\right) \geq \tau \\ e^{-j\frac{4\pi|l-v|}{\lambda_j}} & \cos^{-1}\left(\frac{\boldsymbol{n}\cdot(l-v)}{||\boldsymbol{n}||(l-v)|}\right) < \tau \end{cases} \tag{3}$$

Fig. 6b shows an example when simulating entirely based on specular reflections.

Edge Reflections. In other cases, the signal experiences scattering effects from the edges of objects. First, we define a vertex as being on an edge of an object when the angle between two of its faces is greater than an angle τ_e. We find the set of all vertices that lie on an edge, V_e, and replace Eq. 2 such that it only allows reflections from these vertices:

$$s_j(l,v) = \begin{cases} 0 & v \notin V_e \\ e^{-j\frac{4\pi|l-v|}{\lambda_j}} & v \in V_e \end{cases} \tag{4}$$

Fig. 6c shows an example when simulating edge reflections.

Combining Reflections. In practice, materials rarely exhibit only one type of reflection, but rather some combination of the two. To account for this, our synthetic images can be added together with different weights to simulate different materials:

$$I_{syn}(\alpha_1,\alpha_2) = \frac{\alpha_1}{\alpha_1+\alpha_2} I_s + \frac{\alpha_2}{\alpha_1+\alpha_2} I_e \tag{5}$$

where $I_{syn}(\alpha_1,\alpha_2)$ is the combined synthetic image for weights α_1,α_2, and I_s and I_e are the images from the specular and edge simulations, respectively. Fig. 6d-e shows two examples of combining simulations with different weights representing two different materials.

5.2 Simulation Accuracy

To evaluate the accuracy of our simulator, we use the standard 3D F-score metric [45] to compare the mmWave outputs from real-world and simulation, defined as:

$$PR = \frac{1}{N_R}\sum_{i=1}^{N_R} \mathbb{1}_{d(P_R(i),P_S)<\tau_F}, \quad RE = \frac{1}{N_S}\sum_{j=1}^{N_S} \mathbb{1}_{d(P_S(j),P_R)<\tau_F}, \quad F = \frac{2\,PR\,RE}{PR+RE} \tag{6}$$

where PR, RE, and F are the precision, recall, and F-score. $\mathbb{1}$ is an indicator variable. N_R (N_S) is the number of points in the real (synthetic) point cloud. τ_F is the F-score distance threshold. $d(x,P)$ is the distance from point x to its nearest neighbor in cloud P: $d(x,P) = \min_{x' \in P} ||x - x'||$.

Since our goal is not to select a single set of weights for each object, but to allow the simulation to represent many different objects, we compute this metric across a range of weights $\{\alpha_1,\alpha_2\}$ and choose the combined simulation which produces the best F-score (e.g., the simulation which best matches the properties of this real-world object). The F-score for a synthetic image I_{syn} and real-world image I is $F = \max_{(\alpha_1,\alpha_2)\in W} F(I_{syn}(\alpha_1,\alpha_2),I)$ where W is the set of all weights.

Aligning the Point Clouds. Before we compute the 3D F-score, we first need to convert both the real-world and simulation images to a point cloud. To do so, we apply a standard mmWave point cloud generation procedure[18] by creating one point in the center of each mmWave voxel that has a power above a threshold τ_P.i.e., $P_R = \{p \,||I(p)| > \tau_P\}$ and $P_S = \{p \,||I_{\text{syn}}(p)| > \tau_P\}$, where P_R and P_S are the set of points created for the real and synthetic mmWave images, respectively. Next, we align the real-world and simulation point clouds using a standard iterative closest point implementation [2]. This is necessary since our simulated objects are in the center of the image, while our real world experiments may have objects offset from the center.

Results. We use the above process to compute the 3D F-score for all objects in the dataset. We also compare the accuracy when using only specular simulation images (i.e., $\alpha_1 = 1, \alpha_2 = 0$), and when using only edge simulation images (i.e., $\alpha_1 = 0, \alpha_2 = 1$).

Table 3 reports the 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ percentile F-scores for MITO (1st row), specular only (2nd row) and edge only (3rd row). MITO's simulator achieves a median F-score of 94%, while the specular-only or edge-only simulators achieve median F-scores of 90% and 71%, respectively. An even larger improvement can be seen in the 25$^{\text{th}}$ percentile (85% vs 68% and 49%). This shows the value of MITO's techniques for combining multiple simulation types. An analysis of how simulator accuracy varies with τ_F is in the supplementary.

6 Non-Line-of-Sight Applications

In this section, we evaluate the ability of our dataset and simulator to enable and establish baselines for NLOS perception for diverse, everyday objects. We describe our experimental setup and analyze the evaluation metrics for NLOS segmentation and classification. Such tasks would be useful in a variety of applications, such as confirming customer orders are properly packaged in e-commerce warehouses, detecting broken objects during shipping, and efficiently finding objects for robotic search.

6.1 Object Segmentation

Here, we show how MITO can build a mmWave object segmentation tool that works in LOS and NLOS.
Setup. We build on the state-of-the-art segment-anything model (SAM) [29], which takes as input an image to segment and a prompt, either in the form of points in the image or a bounding box. The model returns one or multiple masks, and a score associated with the predicted mask quality.

Directly using an off-the-shelf object segmentation network for mmWave images faces two challenges. First, SAM is trained on 2D RGB images, but mmWave images are single-channel, complex-valued, and 3D. Second, the SAM model requires a prompt before returning segmentation masks. We note that while there are fully automatic methods using SAM (e.g., iterating through a grid of points as a prompt)[29], they aim to provide all masks within an image, and would not be applicable in our case since they do not identify which mask contains the object of interest.

To convert our mmWave images to 2D RGB images, we start by averaging the magnitudes along the depth of the 3D image, and use a rainbow colormap for colorization, as shown in Fig. 7a. Second, we randomly select a subset of the highest power reflection points as prompts, as shown in Fig. 7b. By feeding the resulting RGB image and prompts to the SAM network, we can obtain mmWave segmentations (Fig. 7c). We apply the same approach for 24 GHz and 77 GHz images. Additional details can be found in the supplementary.

Multi-Spectral Images. Beyond segmenting 77 GHz and 24 GHz images separately, we also propose a method for fusing the two images to improve overall segmentation accuracy. To do so, we combine the

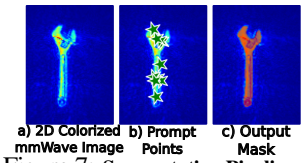


Figure 7: **Segmentation Pipeline.** a) Colorize 2D image, b) Select prompt points, c) Segment with SAM.

Input	LOS					NLOS				
	Pre.	Recall	F1	IoU	Acc.	Pre.	Recall	F1	IoU	Acc.
Camera	**99%**	**99%**	**99%**	**98%**	**99%**	3%	100%*	7%	3%	14%
24 GHz	57%	50%	37%	22%	93%	53%	38%	29%	17%	92%
77 GHz	98%	56%	71%	55%	97%	**95%**	57%	**70%**	**54%**	**97%**
Multi Spectral	97%	62%	74%	59%	98%	**95%**	**63%**	**70%**	**54%**	**97%**

Table 4: **Segmentation Accuracy.** Precision (Pre), Recall, F1 Score, IoU, and Accuracy (Acc) when segmenting camera, 24 GHz, 77 GHz, and multi-spectral images, in both LOS and NLOS scenarios.

images during prompt point selection: instead of using a power threshold based only on a single image, the points must exceed thresholds in both 24 GHz and 77 GHz images. Then, we use these points and the 77 GHz colorized image as input to the SAM network. This allows the 24 GHz image to act as a filter, removing the impact of noise (e.g., from occluders) on the point selection and improving segmentation quality.

6.1.1 Experiments

Baseline. We compare our segmentation performance to an RGB-D camera baseline. We use the camera image from the experiment directly as input to the SAM model. To select prompt points, we use the depth image to automatically filter out pixels greater than a certain depth (i.e., the background), and randomly select prompt points from the remaining pixels. In the LOS scenarios, this corresponds to points on the object. In NLOS scenarios, this corresponds to points on the occluding object (i.e., the cardboard).

Quantitative Results. Tab. 4 shows the median Precision, Recall, F-score, Intersection-over-Union (IoU), and Accuracy when using the camera, 24 GHz, 77 GHz, and multi-spectral (24 GHz & 77 GHz) images for both LOS and NLOS scenarios. In LOS, segmenting camera images achieves a median precision, recall and F-score of 99.7%, 99.5%, and 99.4%, respectively. However, in NLOS, the camera only achieves 3.6% precision and 7.0% F-score, which is expected since cameras (and other visible-light-based sensors) are unable to operate through occlusions. We note that the camera still achieves 100% recall since SAM segments the cardboard occluder, covering the full object.

In contrast, the various mmWave image inputs achieve better performance NLOS scenarios, demonstrating how MITO enables NLOS object segmentation. In particular, the F-score of NLOS for 24 GHz, 77 GHz, and multi-spectral exceeds that of the camera. Notably, multi-spectral images achieve the highest performance. When comparing LOS and NLOS, multi-spectral images only experience a 2.0%, 0.5%, and 3.6% change in precision, recall, and F-score, respectively. This demonstrates that MITO is able to accurately image and perceive fully occluded objects with a similar performance as in LOS.

Interestingly, the recall when segmenting mmWave images is lower (63.1%) than the precision (95.1%). This is primarily because mmWaves are specular, as described earlier, preventing certain portions of the object from appearing in the image. We provide a more detailed discussion on this phenomenon in the supplementary and Sec. 6.3 highlight how future work can improve recall via shape completion.

The IoU and accuracy show similar trends. Notably, 24 GHz, 77 GHz, and multi-spectral all outperform the camera in NLOS. We note that the accuracy is high because it is dominated by the background.

Qualitative Results. In some cases, the depth camera is unable to measure the depth to every point on the object. This is typically the case for very reflective objects (e.g., mirror-like metallic objects) or clear objects (e.g., glass or plastic). In these cases, MITO is able to reconstruct more of the object than the depth camera, as shown in Fig. 8. In Fig. 8b, the image from the depth camera has holes due to the object's material reflectivity, whereas in Fig. 8c, MITO's mmWave image captures the whole object. Fig. 8d shows our 77 GHz mmWave segmentation masks the entire object.

6.2 Sim2Real Shape Classification

Next, we demonstrate how MITO can be used to classify objects in NLOS based on their geometry.

Setup. We design a classifier which takes 77 GHz mmWave images (after applying the segmentation mask from Sec. 6.1) as input, and outputs the estimated class label. The classifier is a neural network consisting of 5 convolutional layers, followed by 3 fully connected layers, and a softmax activation. We *train the classifier entirely on synthetic data*, and *evaluate on real-world* images in both LOS and NLOS. To generate synthetic images for training, we leverage object meshes provided by YCB dataset.

One challenge in training from synthetic images is to determine which types of reflections dominate for a given object. Recall from Sec. 5 that our simulator models two different reflection types. However, it

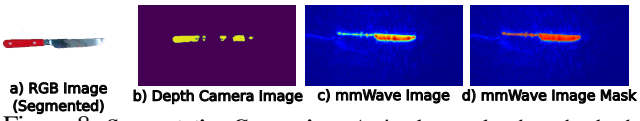


Figure 8: **Segmentation Comparison.** A visual example where the depth image is unable to capture the entire object, but the mmWave image does.

Training Data		Accuracy		
Edge	**Spec**	**LOS**	**NLOS**	**Overall**
X		61.1%	75.0%	67.6%
	X	77.7%	75.0%	76.4%
X	X	**88.8%**	**81.2%**	**85.2%**

Table 5: **Classifier Accuracy** for LOS, NLOS, and overall using specular (Spec), edge, and combined reflection simulations for training.

is difficult to accurately measure the material/texture properties of all objects to identify how to correctly combine these simulation images. Instead, we introduce a new data augmentation, where we choose random weights for each simulation type and sum the weighted images. This allows the network to train on different material properties for each object. We also apply standard image augmentations (rotation, translation, masking, blur). More training details are in supplementary.

Experiments. We report the shape classification results on real-world data in LOS and NLOS in Tab. 5. Our classifier performs with an overall accuracy of 85.2%. Specifically, it achieves an accuracy of 88.8% and 81.2% for LOS and NLOS scenarios, respectively, showing that MITO can be used to enable novel capabilities such as NLOS shape classification.

In Tab. 5, we also include results using a classifier trained on either only edge images (i.e., without combining them with the specular images) or only specular images in order to evaluate the impact of our different simulation types on the classifier accuracy. We observe that these models are only able to achieve 67.6% and 76.4%, whereas using both simulation types together achieves 85.2% accuracy. This shows the benefit of our simulator's different outputs, and our data augmentation technique to combine them.

These results shows that MITO has the potential to be used for novel NLOS applications, such as confirming packaged orders in closed boxes. Furthermore, by training on synthetic images and evaluating on real-world data, we show that our simulator is representative of the real-world.

6.3 Additional NLOS Applications & Extensions

In addition to the above applications, MITO also paves the way for future work in multiple directions:

- *NLOS Object Completion:* In some cases, mmWave images do not capture the entire object due to specularity (see supplementary for more details). It would be interesting future work to create an object completion network that learns the missing parts of an object and creates more photo-realistic NLOS images.
- *NLOS 6DoF Estimation:* Another interesting problem is 6DoF estimation from mmWave images. This could be used to ensure proper liquid packaging to avoid spills, or enable efficient grasp planning in NLOS (e.g., grasping an item under cloth or within a box with packing peanuts).
- *NLOS Robotic Manipulation:* Future work can leverage many of the above NLOS perception tasks to enable new tasks in robotic manipulation. For example, accurate mmWave object segmentation and classification can significantly increase the efficiency of mechanical search and retrieval (i.e., the problem of searching for a specific item in clutter). Further, mmWave images can be used to analyze pile stability to improve push and grasp planners. NLOS perception can also help autonomous robots determine which packages contain fragile materials to be handled with special care. More generally, NLOS perception can help autonomous robots to operate in more realistic, cluttered environments.

7 Conclusion

We present MITO, the first millimeter-wave (mmWave) dataset of diverse, everyday objects in both line-of-sight and fully-occluded settings. We collect over 24 million frames of mmWave images, and used them to synthesize 550 high-resolution real world 3D (SAR) mmWave images for 76 objects. We also develop an open-source simulator that can be used to generate synthetic images for any 3D mesh. We demonstrate the utility of our contributions in two applications – NLOS object segmentation and NLOS shape classification – establishing baselines for future research and paving the way for important applications in robotic perception, logistics, scene understanding.

As the research evolves, it would be interesting to expand the dataset (e.g., incorporating additional items, surfaces, occlusion types, etc.) which would continue improving the accuracy of the demonstrated NLOS tasks and unlock additional ones. More generally, we hope this work serves as a foundation and motivation for future research in mmWave-based NLOS perception.

References

[1] Ti dca1000evm. https://www.ti.com/tool/DCA1000EVM. 4

[2] Icp registration. https://www.open3d.org/docs/release/tutorial/pipelines/icp_registration.html. 7

[3] Perceiving systems mesh package. https://ps.is.mpg.de/code/mesh-library. 6

[4] Infineon position2go. https://www.infineon.com/cms/en/product/evaluation-boards/demo-position2go/. 4

[5] Understanding mmwave for 5g networks. https://www.5gamericas.org/wp-content/uploads/2020/12/InDesign-Understanding-mmWave-for-5G-Networks.pdf. 4

[6] Ti iwr1443boost. https://www.ti.com/product/IWR1443#tech-docs. 1, 4

[7] mmWave Radar Radome Design Guide. https://www.ti.com/lit/an/swra705/swra705.pdf, 2021. Texas Instruments. 4

[8] Ansys high frequency structure simulator (hfss), 2022. Ansys HFSS. 2, 3, 5

[9] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C Miller. 3d tracking via body radio reflections. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*, pages 317–329, 2014. 4

[10] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)*, 34(6):219, 2015. 4

[11] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[12] Walter Brescia, Pedro Gomes, Laura Toni, Saverio Mascolo, and Luca De Cicco. Millinoise: a millimeter-wave radar sparse point cloud dataset in indoor scenarios. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 422–428, 2024. 2

[13] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.01164. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.01164. 2, 3

[14] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017. doi: 10.1177/0278364917700714. URL https://doi.org/10.1177/0278364917700714. 2, 4, 5

[15] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar, 09 2022. 2, 3

[16] Han Cui, Shu Zhong, Jiacheng Wu, Zichao Shen, Naim Dahnoun, and Yiren Zhao. Milipoint: A point cloud dataset for mmwave radar, 2023. 2, 3

[17] Pham Tien Dat, Yuya Yamaguchi, Shingo Takano, Shotaro Hirata, Junichiro Ichikawa, Ryo Shimizu, Keizo Inagaki, Isao Morohashi, Yuki Yoshida, Atsushi Kanno, Naokatsu Yamamoto, Tetsuya Kawanishi, and Kouichi Akahane. Terahertz signal transparent relay and switching using photonic technology. *Journal of Lightwave Technology*, 42(3):1173–1182, 2024. doi: 10.1109/JLT.2023.3325277. 4

[18] Laura Dodds, Hailan Shanbhag, Junfeng Guan, Saurabh Gupta, and Haitham Hassanieh. Around the corner mmwave imaging in practical environments. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2024. 7

[19] Sung Yun Jun et. al. Penetration loss at 60 ghz for indoor-to-indoor and outdoor-to-indoor mobile scenarios. In *14th EuCAP*, 2020. doi: 10.23919/EuCAP48036.2020.9135581. 14

[20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2

[21] Federal Communications Commission. Frequency Allocation:. transition.fcc.gov/oet/spectrum/table/fcctable.pdf. 4

[22] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1737–1746, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310154. doi: 10.1145/2207676.2208303. URL https://doi.org/10.1145/2207676.2208303. 2

[23] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through fog high-resolution imaging using millimeter wave radar. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[24] Runwei Guan, Ruixiao Zhang, Ningwei Ouyang, Jianan Liu, Ka Lok Man, Xiaohao Cai, Ming Xu, Jeremy Smith, Eng Gee Lim, Yutao Yue, et al. Talk2radar: Bridging natural language with 4d mmwave radar for 3d referring expression comprehension. *arXiv preprint arXiv:2405.12821*, 2024. 2, 3

[25] Siddharth Gupta, Prabhat Kumar Rai, Abhinav Kumar, Phaneendra K. Yalavarthy, and Linga Reddy Cenkeramaddi. Target classification by mmwave fmcw radars using machine learning on range-angle images. *IEEE Sensors Journal*, 21(18):19993–20001, 2021. doi: 10.1109/JSEN.2021.3092583. 3

[26] Jakob Hoydis, Fayçal Aït Aoudia, Sebastian Cammerer, Merlin Nimier-David, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna rt: Differentiable ray tracing for radio propagation modeling. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 317–321. IEEE, 2023. 2, 5

[27] Intel RealSense. https://www.intelrealsense.com, 2019. 4

[28] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T. Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1168–1174, 2011. doi: 10.1109/ICCVW.2011.6130382. 2

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 5, 7

[30] Andrew Kramer, Kyle Harlow, Christopher Williams, and Christoffer Heckman. Coloradar: The direct 3d millimeter wave radar dataset. *The International Journal of Robotics Research*, 41(4):351–360, 2022. doi: 10.1177/02783649211068535. URL https://doi.org/10.1177/02783649211068535. 2, 3

[31] Mohamed Lamane, Mohamed Tabaa, and Abdessamad Klilou. Classification of targets detected by mmwave radar using yolov5. *Procedia Computer Science*, 203:426–431, 2022. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.07.056. URL https://www.sciencedirect.com/science/article/pii/S1877050922006615. 17th International Conference on Future Networks and Communications / 19th International Conference on Mobile Systems and Pervasive Computing / 12th International Conference on Sustainable Energy Information Technology (FNC/MobiSPC/SEIT 2022), August 9-11, 2022, Niagara Falls, Ontario, Canada. 3

[32] Jie Liu, Kai Zhang, Zhenlin Sun, Qiang Wu, Wei He, and Hao Wang. Concealed object detection and recognition system based on millimeter wave fmcw radar. *Applied Sciences*, 11(19), 2021. ISSN 2076-3417. doi: 10.3390/app11198926. URL https://www.mdpi.com/2076-3417/11/19/8926. 3

[33] Ting Liu, Yao Zhao, Yunchao Wei, Yufeng Zhao, and Shikui Wei. Concealed object detection for activate millimeter wave image. *IEEE Transactions on Industrial Electronics*, 66(12):9909–9917, 2019. doi: 10.1109/TIE.2019.2893843. 2, 3

[34] Jonathan S. Lu, Patrick Cabrol, Daniel Steinbach, and Ravikumar V. Pragada. Measurement and characterization of various outdoor 60 ghz diffracted and scattered paths. In *MILCOM 2013 - 2013 IEEE Military Communications Conference*, pages 1238–1243, 2013. doi: 10.1109/MILCOM.2013.212. 4

[35] Marco Mezzavilla, Menglei Zhang, Michele Polese, Russell Ford, Sourjya Dutta, Sundeep Rangan, and Michele Zorzi. End-to-end simulation of 5g mmwave networks. *IEEE Communications Surveys & Tutorials*, 20(3):2237–2263, 2018. 2

[36] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=W_bsDmzwaZ7. 2

[37] Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila. Multi-class road user detection with 3+1d radar in the view-of-delft dataset. *IEEE Robotics and Automation Letters*, 7(2):4961–4968, 2022. doi: 10.1109/LRA.2022.3147324. 2, 3

[38] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. Radarhd: Demonstrating lidar-like point clouds from mmwave radar. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–3, 2023. 2

[39] M Mahbubur Rahman, Ryoma Yataka, Sorachi Kato, Pu Perry Wang, Peizhao Li, Adriano Cardace, and Petros Boufounos. Mmvr: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. *arXiv preprint arXiv:2406.10708*, 2024. 2, 3

[40] Hem Regmi, Moh Sabbir Saadat, Sanjib Sur, and Srihari Nelakuditi. Squigglemilli: Approximating sar imaging on mobile millimeter-wave devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(3), sep 2021. doi: 10.1145/3478113. URL https://doi.org/10.1145/3478113. 2, 3, 5

[41] Bofeng Su and Minghui Yuan. Object recognition for millimeter wave mimo-sar images based on high-resolution feature recursive alignment fusion network. *IEEE Sensors Journal*, 23(14): 16413–16427, 2023. doi: 10.1109/JSEN.2023.3284480. 2, 3

[42] Peng Sun, Ting Liu, Xiaotong Chen, Shiyin Zhang, Yao Zhao, and Shikui Wei. Multi-source aggregation transformer for concealed object detection in millimeter-wave images. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6148–6159, 2022. doi: 10.1109/TCSVT.2022.3161815. 3

[43] Shu Sun, George R MacCartney, and Theodore S Rappaport. A novel millimeter-wave channel simulator and applications for 5g wireless communications. In *2017 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2017. 2

[44] Y. Sun, Z. Huang, H. Zhang, Z. Cao, and D. Xu. 3drimr: 3d reconstruction and imaging via mmwave radar based on deep learning. In *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, pages 1–8, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/IPCCC51483.2021.9679394. URL https://doi.ieeecomputersociety.org/10.1109/IPCCC51483.2021.9679394. 3

[45] Maxim Tatarchenko*, Stephan R. Richter*, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? 2019. 6

[46] Universal Robots, UR5e. https://www.universal-robots.com/products/ur5-robot/, 2021. 4

[47] Yizhou Wang, Jen-Hao Cheng, Jui-Te Huang, Sheng-Yao Kuan, Qiqian Fu, Chiming Ni, Shengyu Hao, Gaoang Wang, Guanbin Xing, Hui Liu, et al. Vision meets mmwave radar: 3d object perception benchmark for autonomous driving. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2769–2775. IEEE, 2024. 2

[48] Zhongce Wang, Chao Jin, Lingzhang Sun, and Yunfan Li. Wrdi: A multimodal dataset of mmwave radar data and image. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3209–3214, 2021. doi: 10.1109/BigData52589.2021.9671545. 2, 3

[49] Sahil Waqar and Matthias Pätzold. A simulation-based framework for the design of human activity recognition systems using radar sensors. *IEEE Internet of Things Journal*, 2023. 2, 5

[50] Yingxiao Wu, Zhongmin Jiang, Haocheng Ni, Changlin Mao, Zhiyuan Zhou, Wenxiang Wang, and Jianping Han. mmhpe: Robust multi-scale 3d human pose estimation using a single mmwave radar. *IEEE Internet of Things Journal*, 2024. 2, 3

[51] Jianjian Yang, Tianmu Gui, Yibo Tong, Yuyuan Zhang, Qiankun Huang, and Guanghui Zhao. Autonomous mining transportation systems: Integrating 4d mmwave radar for enhanced detection of obstructed static objects. *IEEE Transactions on Intelligent Vehicles*, 2024. 2, 3

[52] Xiaotong Zhang, Zhenjiang Li, and Jin Zhang. Synthesized millimeter-waves for human motion sensing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, page 377–390, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398862. doi: 10.1145/3560905.3568542. URL https://doi.org/10.1145/3560905.3568542. 2, 5

[53] Lianqing Zheng, Zhixiong Ma, Xichan Zhu, Bin Tan, Sen Li, Kai Long, Weiqi Sun, Sihan Chen, Lu Zhang, Mengyue Wan, Libo Huang, and Jie Bai. Tj4dradset: A 4d radar dataset for autonomous driving. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 493–498, 2022. doi: 10.1109/ITSC55140.2022.9922539. 2, 3

[54] Junyu Zhou and Julien Le Kernec. 4d radar simulator for human activity recognition. *IET Radar, Sonar & Navigation*, 2023. 2, 5

A Code and Dataset

The code and dataset are released here: https://github.com/signalkinetics/MITO_Codebase

B Object Diversity

Since our dataset represents the vast majority of the YCB dataset, our objects are diverse across shape, size, material, and type. In terms of shape, our dataset includes spherical, flat, curved, and complex shapes. In terms of size, it includes objects whose dimensions range from 1.5cm to 47cm. Our objects include various materials such as wood, plastic (e.g., HDPE, PETE, Polypropelene, Nylon, etc), metal, cardboard, rubber, as well as objects with multiple materials. Additionally, our objects represent a wide range of categories, including household tools (wrench, screwdriver, hammer, etc.), groceries (sugar box, soup can, cracker box, etc.), kitchen supplies (spatula, plate, skillet, etc.), toys (marbles, rubik's cube, baseball, etc.), and office supplies (scissors, dry erase markers, etc.).

C Why other NLOS imaging techniques are not applicable?

Alternative NLOS sensing technologies face significant limitations for our scenario. Ultrasound cannot image through closed boxes due to the high acoustic impedance mismatch between air and common occlusion materials (e.g., cardboard). In medical applications, gels are used on Ultrasound transducers to reduce this mismatch with human tissue, which is not practical for object imaging. X-ray signals, while capable of penetrating occlusions, involve ionizing radiation and are therefore unsafe for extended human exposure—unlike mmWave signals, which are non-ionizing and comparable to Wi-Fi. Finally, thermal imaging detects heat sources in NLOS conditions, but this capability does not apply here since the objects of interest do not emit significant heat.

D Sample Dataset Images

Here, we include sample mmWave images for 50 out of the 76 objects in MITO. Note that the majority of our objects contain multiple real-world images from different angles or measurments apertures.

Fig. 13 shows the segmented RGB images (1st column), 77GHz line-of-sight (2nd column), 77GHz non-line-of-sight (3rd column), 24GHz line-of-sight (4th column), and 24GHz non-line-of-sight (5th

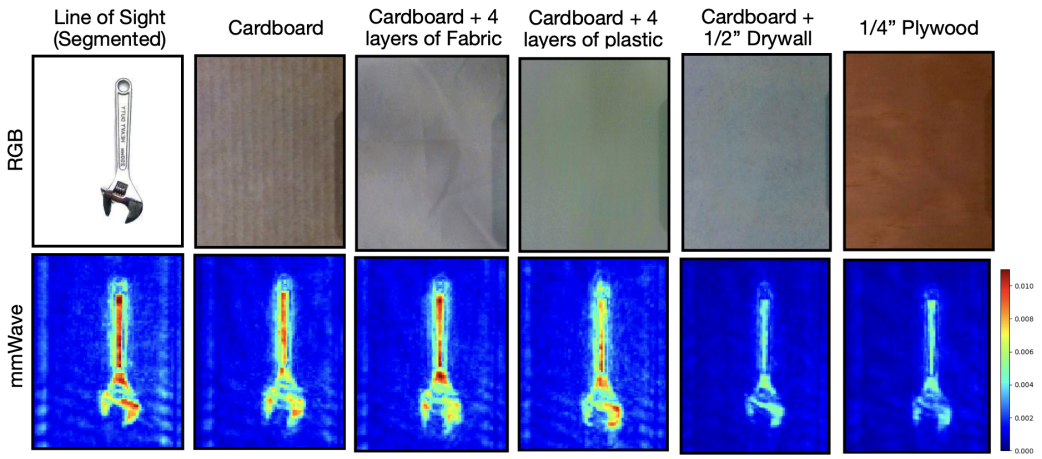


Figure 9: Occlusions Materials impact on mmWave

column) mmWave images from real-world experiments. The objects are sorted by their unique identifier (defined by the YCB dataset). This shows the diversity of objects within MITO, as well as the performance of the mmWave imaging for different objects.

E Impact of Occlusion types

Different occlusion materials can influence the mmWave signals propagation and thus the quality of mmWave images. Prior wireless research has studied how different materials attenuate RF signals [19], often modeling this effect via changes in signal-to-noise ratio. Thus, we expect different occlusions to primarily change the brightness or noisiness of our mmWave images.[4] We verify this through a qualitative analysis in Fig. 9, which shows the RGB and mmWave images of a metallic wrench. It compares LOS to multiple occlusion types (cardboard, fabric, plastic, drywall, & plywood), plotting all heatmaps with the same color scale.[5] We note that:

- The wrench is clearly visible in all images, showing that mmWave imaging works through various occlusions.
- The intensity of the wrench is lower for the drywall and wood (note that the images do not contain any red regions), which is expected since these materials introduce greater attenuation.
- Some occlusions show increased noise around the wrench, especially for the deformable plastic and fabric occlusions. These occlusions do not lie completely flat, and thus can induce small distortions in the image. This highlights the need for models and algorithms which can account for mmWave-specific behavior to enable NLOS perception tasks. Such occlusions would be valuable additions for future extensions of MITO.

F Aligning RGB-D and mmWave Data

Many perception tasks require ground-truth knowledge of an object's shape and location for training and evaluation. For example, our segmentation task in Sec. 6.1 of the main paper requires ground-truth 2D binary masks of the object for evaluation. Other tasks, such as surface reconstruction, 6D pose estimation, and 3D segmentation, would require a ground-truth 3D point cloud of the object's surface.

In this section, we describe how we obtain ground-truth object point clouds and segmentation masks from RGB-D data, and align these with the 3D mmWave image. We leverage RGB-D images captured from a wrist mounted camera at the beginning of each experiment (when the object is still in line-of-sight). First, to isolate the object within the RGB-D image, we manually segment the object within the RGB image (see Sec. 4.3 in the main paper for more details). Then, we apply the following four steps to align the

[4] Metal obstacles may block all wireless signals and prevent imaging.

[5] Our occlusions are 1 layer of cardboard, 1 layer of cardboard plus 4 layers of a fabric sheet, 1 layer of cardboard plus four layers of a plastic shower curtain, 1 layer of cardboard plus 1/2" drywall, and 1 sheet of 1/4" plywood.

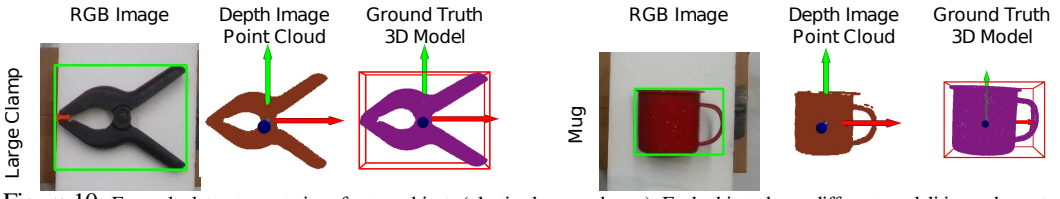


Figure 10: Example dataset annotations for two objects (plastic clamp and mug). Each object shows different modalities and annotation types: RGB image with 2D bounding box, depth image point cloud, and ground truth 3D mesh with 6-DoF pose and bounding box.

segmented RGB-D data with the mmWave image and produce both ground-truth 2D segmentation masks and 3D point clouds:

1. Transforming mmWave data: First, we use the radar's poses during the mmWave image acquisition to transform the 3D mmWave images to the world frame:

$$^{W}P^{U} = {}^{W}R^{M} \times {}^{M}P^{U} + {}^{W}P^{M} \tag{7}$$

where $^{W}P^{U}$ and $^{M}P^{U}$ are the position (x,y,z) of a voxel U in the mmWave image in the world frame W and mmWave radar frame M, respectively; $^{W}P^{M}$ and $^{W}R^{M}$ are the position and rotation of the radar frame M in the world frame.

2. Transforming Visual Data: Next, we use the RGB-D data to produce a 3D point cloud of the object in the world frame. To do so, we apply the manually segmented RGB mask to the camera's depth image.[6] Then, we convert each point within the segmented depth image into a 3D point using the camera's intrinsic properties. Finally, we use the camera's pose to convert this point cloud from the camera's frame to the world frame:

$$^{W}P^{V} = {}^{W}R^{C} \times {}^{C}P^{V} + {}^{W}P^{C} \tag{8}$$

where $^{W}P^{V}$ and $^{C}P^{V}$ are a point (x,y,z) from vision-based point cloud in the world frame W and camera frame C, respectively; $^{W}P^{C}$ and $^{W}R^{C}$ are the position and rotation of the camera frame in the world frame, respectively.

This point cloud is now in the same frame as the mmWave image in the world frame ($^{W}P^{U}$ in Eq. 7) and can serve as 3D ground truth for different perception tasks.

3. Constructing 2D Binary Masks: Next, we produce ground-truth binary masks. To do so, we voxelize the vision-based point cloud ($^{W}P^{V}$ in Eq. 8). Here, each voxel's center and size is the same as that of the mmWave image's voxels. Then, we project both the 3D vision-based image and 3D mmWave image into two dimensions. Now, the projected 2D vision-based image is a binary mask which acts as a ground-truth segmentation for the 2D mmWave image. It can also be used to generate a 2D bounding box of the object as shown in the RGB images in Fig. 10.

4. Creating 3D Ground-Truth Models. We also provide ground-truth 3D models and 3D bounding boxes. To do so, we leverage meshes from the YCB dataset for each object, and convert them into point clouds by uniformly sampling points along the surface. Then, we align these point clouds to the vision-based point clouds in the world frame ($^{W}P^{V}$). We use the iterative closest point (ICP) algorithm to find the transformation and rotation which best aligns the full 3D point cloud to the measured depth information. This provides a ground-truth 3D representation of the object in the world frame, which can also be used to derive a 3D bounding box [7] We show examples of this annotation in Fig. 10.

G Radar & SAR Details:

Our radars leverage frequency-modulated continuous wave (FMCW) signals, which span a large bandwidth. The bandwidths are 3 GHz for the 77 GHz radar and 200 MHz for the 24 GHz radars, respectively.

Given the bandwidth and aperture, we can compute MITO's resulting resolution. MITO's resolution in the X,Y, and Z dimensions (computed using Eqs. 2-3 in the main paper) are 3.1 mm, 1.5 mm, and 4.9 cm for the 77 GHz radar and 1 cm, 0.5 cm, and 0.75 m for the 24 GHz radar, respectively.

[6] The realsense depth camera provides aligned RGB and depth images.
[7] We find this bounding box using Open3D's `get_axis_aligned_bounding_box` function.

H Task 1: Object Segmentation

Here, we provide additional details on the object segmentation from Sec 6.1 in the main paper.

H.1 Additional Design Details

We describe how we leverage SAM for mmWave segmentation in more detail. To do so, we first need to convert our 3D, complex-valued, single-channel mmWave images to 2D RGB images that can be input to the SAM model. We start by projecting our 3D image into a real, 2D image by averaging the magnitudes along the depth of the image. Then, as illustrated in the example in Fig. 7a in the main paper, we colorize the single-channel image using a rainbow colormap, such that high values appear red and low values appear blue. The resulting three-channel image can then be input into SAM.

Next, we need to choose one or more points to prompt the network. One may think that we can simply choose the highest value pixel as the prompt. However, this typically results in an insufficient mask that does cover the entire object. Instead, we choose a number of points that are dispersed throughout the object such that the network is encouraged to segment the entire object. To do so, we first find the set of all points within a fixed percentage of the maximum power, S_p:

$$S_p = \{p \mid |I(p)| > \tau * \max(|I|)\} \tag{9}$$

where $|I(p)|$ is the magnitude of the mmWave image at point p, τ is a threshold percentage.

Since this will be a large number of points, we then randomly choose a subset:

$$P_p = \{p_i \mid p_i \in_R S_p \cap i \in \{1,..,K\}\} \tag{10}$$

where P_p is the final set of points used to prompt the SAM network, K is the number of chosen points, and $\in_R$ denotes a randomly selecting from a set. An example of these points is shown in Fig. 7b.

Now, the converted RGB image and prompt points are fed to the SAM network to produce mmWave segmentations. An example output is shown in Fig. 7c by a red mask over the segmented pixels.

H.2 Implementation Details

We implement the SAM model using pre-trained weights provided by Meta. Then, SAM can be used to output multiple possible segmentation masks for a given prompt, along with a score for each mask. To remove the invalid mmWave masks, we remove any masks that take up a large percentage of the image. From the remaining masks, we select the mask with the highest score.

H.3 Evaluation Details

Evaluation Metrics: We evaluate our segmentation with precision, recall, F-Score, IoU, and Accuracy. These are defined as:

$$P = \frac{\texttt{TP}}{\texttt{TP}+\texttt{FP}} \qquad R = \frac{\texttt{TP}}{\texttt{TP}+\texttt{FN}} \tag{11}$$

$$F = \frac{2\,P\,R}{P+R} \tag{12}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{13}$$

$$Acc = \frac{TP+TN}{TP+TN+FN+FP} \tag{14}$$

where P, R, F, IoU, and Acc are the precision, recall, F-Score, IoU, and Accuracy, respectively. TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively.

Ground Truth: To evaluate the object segmentation, we use the ground truth masks described in Sec. F.3. Note that the depth image captured by the camera sometimes contains NaNs due to its inability to accurately capture some parts of objects. To resolve this problem, we fill in any missing depth value with the average depth value of the mask before transforming it to the radar frame.

Evaluation: Since our point selections relies on a random process, we measure the above metrics 3 times for each object using a different instance of prompt selection and segmentation. We evaluate our segmentation on a subset of our dataset across 236 real-world images (both 24 GHz and 77 GHz).

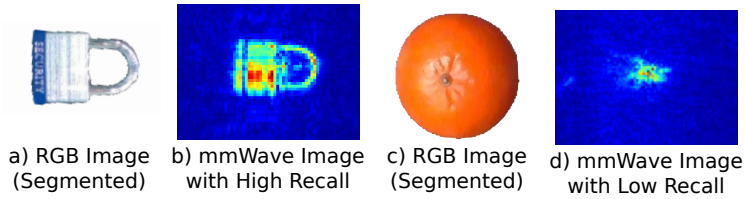


Figure 11: **Recall.** This figure shows two examples where the mmWave image exhibits high and low recall. a) shows the segmented RGB image for a padlock, and b) shows its corresponding mmWave image, where the full object is captured in the image, and results in high recall. c) shows a second example of the segmented RGB image for a spherical orange. d) is the corresponding mmWave image for this object. Due to specular reflection of mmWave from the orange's surface, only a small portion of this object is captured in the mmWave image.

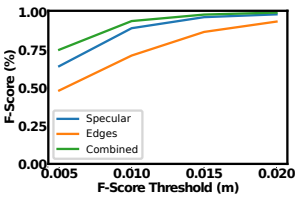


Figure 12: Impact of F-Score threshold on simulator accuracy

H.4 Analyzing Segmentation Performance

As shown in Sec. 6.1 in the main paper, the median recall of our segmentation (63%) is lower than the precision (95%). To understand why, we analyze an example of two different objects in Fig. 11. In the first example, Fig. 11a shows the RGB image of a padlock. For this object, Fig. 11b shows that the mmWave image captures a large portion of the object. This is because the surface is fairly flat.

In the second example, Fig. 11c shows an example of a spherical orange. Fig. 11d shows the mmWave image for this object. In this case, the object is not fully captured in the mmWave image. This is because, at millimeter-wave frequencies, the object is experiencing specular reflections (see Sec. 3.3 in the main paper). Since the object is spherical, the sides of the object will reflect signals away from the radar, preventing the radar from receiving them and the mmWave imaging algorithm from recovering the full object's shape. This leads to a low segmentation recall performance when compared with the ground-truth segmentation mask produced from the RGB image.

In contrast to mmWave images, visible light experiences primarily diffuse reflections from the surface of the orange, which results in full recovery of the orange's shape within the camera image. This is one example of how mmWaves images are fundamentally different from camera images. Further, it shows the potential benefit for future work on generalizable mmWave object completion.

I Task 2: Sim2Real Shape Classification

Here, we provide additional details on the shape classification from Sec. 6.2 in the main paper.

I.1 Implementation Details

Network Design: We design a neural network with 5 convolutional layers, followed by 3 fully connected layers, and a softmax activation. The input to the network is a mmWave image (after applying the binary mask generated in Sec. 6.1). The network then outputs the class probabilities. We sort our dataset into 4 classes: Boxes, Cans, Spheres, and Handle-based tools[8].

Training: We train our network on an Ubuntu 22.04 computer with a GTX 1080 GPU. We use an adam optimizer with a learning rate of 1e-3, and we train our network for 500 epochs. Since our classes are unbalanced, we sample our data during training such that each class is equally represented.

Augmentations: In addition to our mmWave augmentation described in Sec. 6.2, we also implement a number of other randomized augmentations: rotations, translations, scaling, masking, horizontal and vertical flipping.

Data Split: The training set is comprised entirely of simulation data. Then, the real world data (LOS and NLOS) is randomly distributed into a test and validation set (with equal sizes), such that each class is represented equally in the test and validation set.

[8]For the purpose of this classifier, we do not use objects from the dataset that do not fit in these categories when training/testing the classifier.

I.2 Evaluation Details

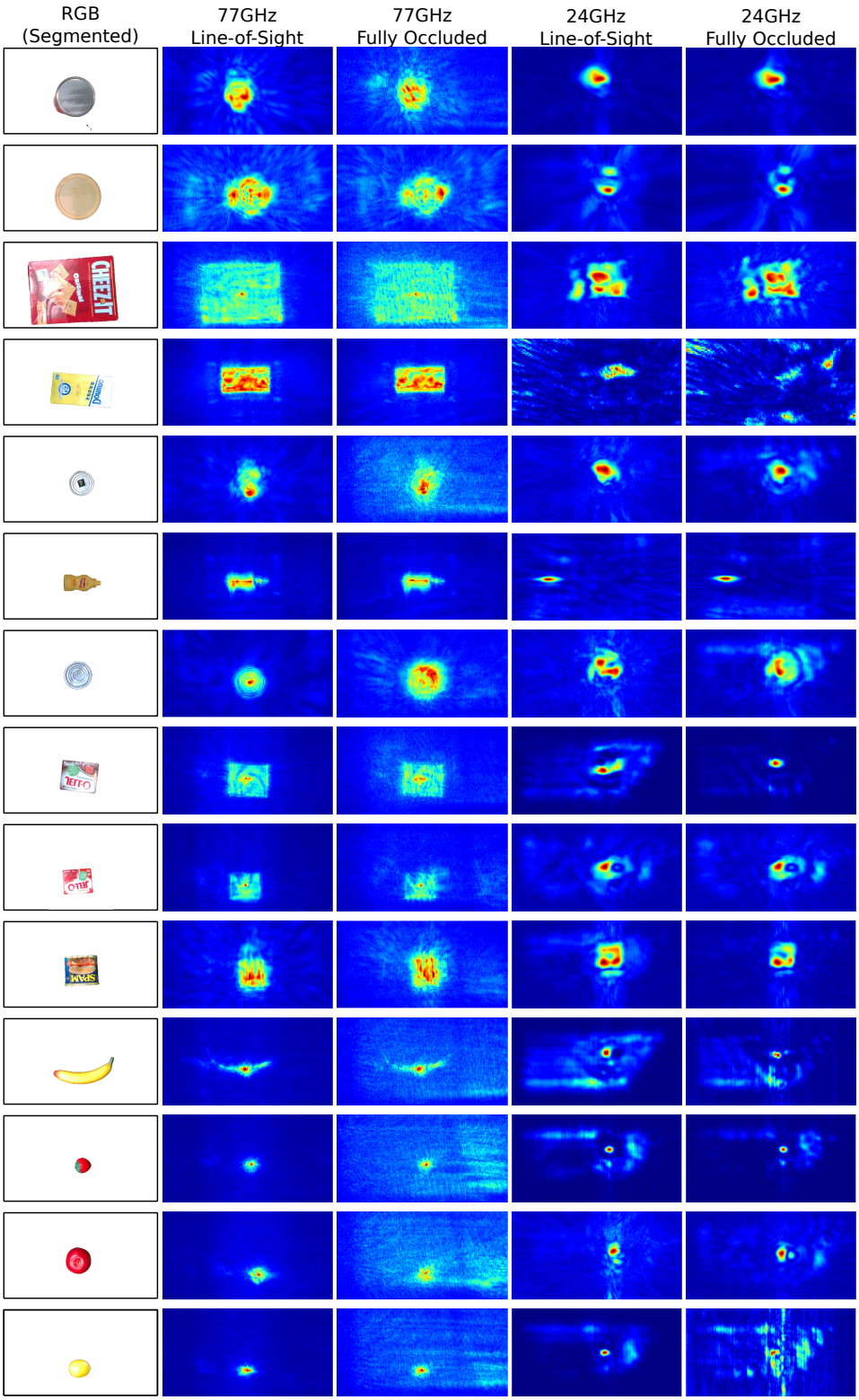
Evaluation Metrics: We evaluate our mmWave classification with the accuracy, defined by the percentage of correctly classified images in the test set.

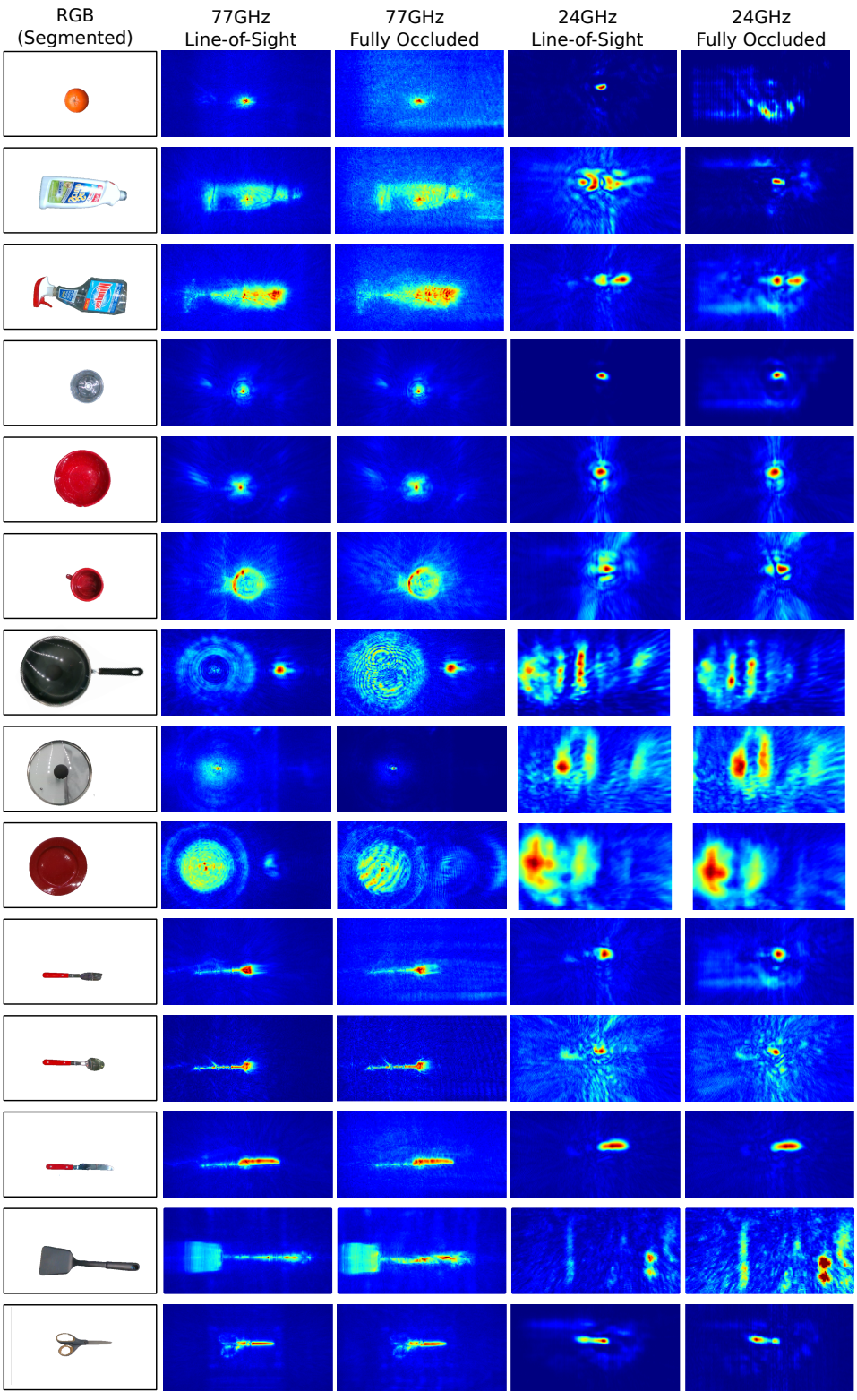
Ground Truth: For the object classifier, we use the ground truth labels for each experiment and manually sort the objects into the above classes.

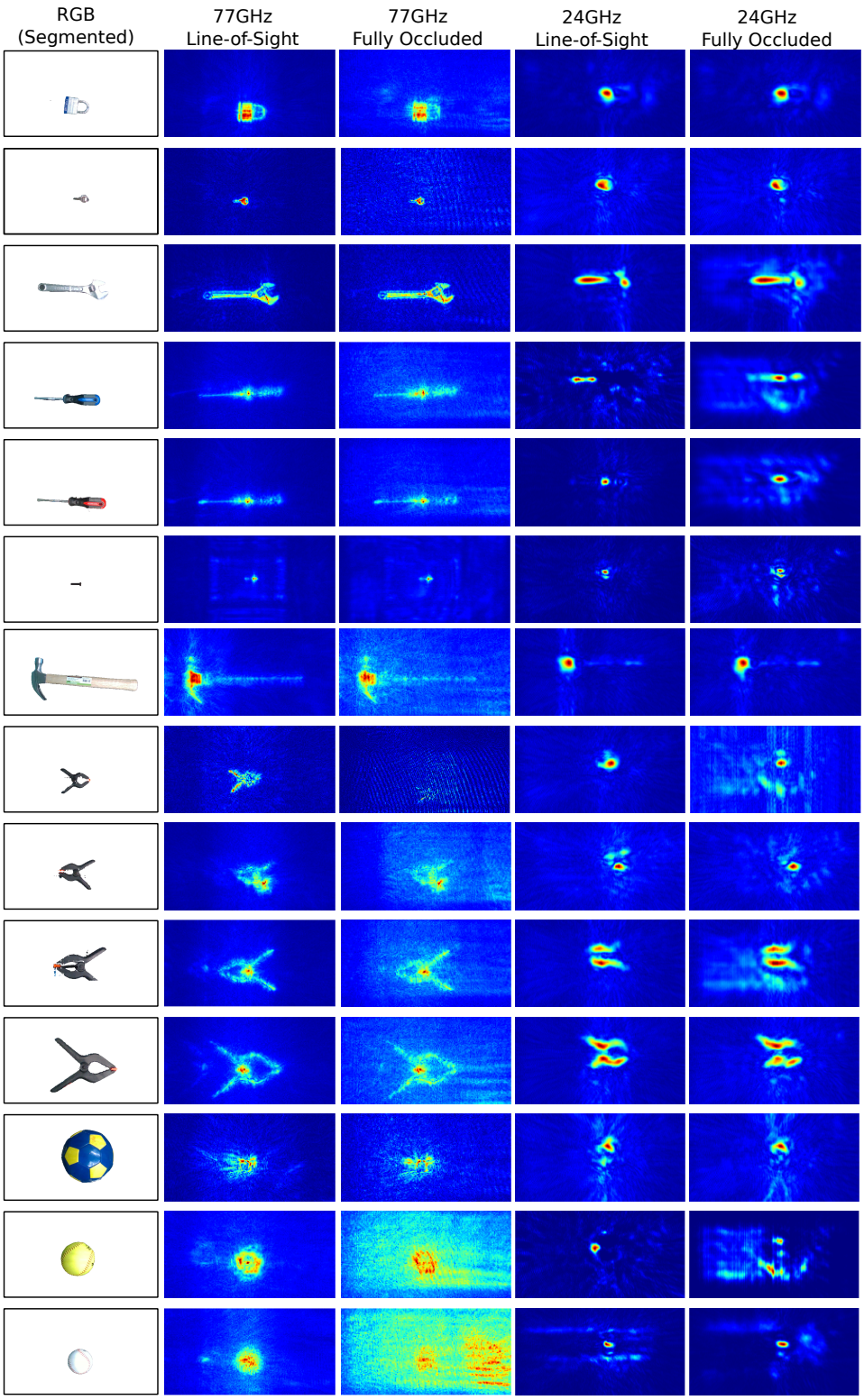
J Impact of F-Score threshold on simulator accuracy

In the main paper (Sec 5.2), simulator accuracy is evaluated using the 3D F-score metric with a fixed threshold. This threshold, denoted as τ_F in Equ. 9 and 10, determines the maximum allowable distance between corresponding points in the real and synthetic point clouds for them to be considered a match. To better understand the sensitivity of simulator performance to this parameter, we analyze how the F-score varies across a range of thresholds from 0.005 m to 0.02 m. Fig. 12 plots the F-score for all objects in the dataset under three simulation settings: combined simulation (green), specular-only (blue), and edge-only (orange).

We observe that even at a strict threshold of 0.005 m, the combined simulation maintains an F-score above 75%, indicating strong alignment with real-world data. As the threshold increases, all methods show improved scores, but the combined simulation consistently outperforms the others. We also note that F-Score thresholds $\geq$1 cm are common for mmWave-based systems [18], as they have lower resolutions than their RGB-D counterparts.


RGB
(Segmented)
77GHz
Line-of-Sight
77GHz
Fully Occluded
24GHz
Line-of-Sight
24GHz
Fully Occluded


RGB
(Segmented)
77GHz
Line-of-Sight
77GHz
Fully Occluded
24GHz
Line-of-Sight
24GHz
Fully Occluded


RGB
(Segmented)
77GHz
Line-of-Sight
77GHz
Fully Occluded
24GHz
Line-of-Sight
24GHz
Fully Occluded

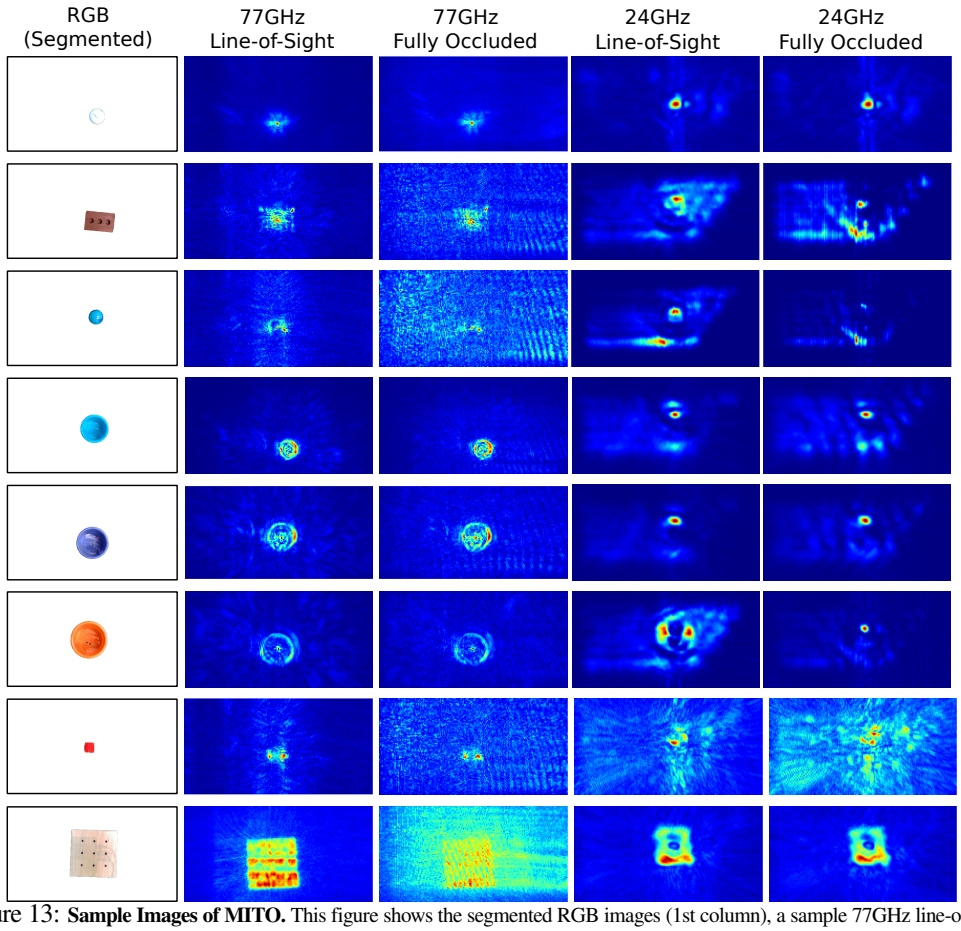


Figure 13: **Sample Images of MITO.** This figure shows the segmented RGB images (1st column), a sample 77GHz line-of-sight (2nd column) 77GHz non-line-of-sight (3rd column), 24GHz line-of-sight (4th column) and 24GHz non-line-of-sight (5th column) mmWave image from a real-world experiment.