Exploring mmWave Radar and Camera Fusion for High-Resolution and Long-Range Depth Imaging

Akarsh Prabhakara, Diana Zhang, Chao Li, Sirajum Munir, Aswin C. Sankaranarayanan,
Anthony Rowe and Swarun Kumar

***Abstract*— Robotic geo-fencing and surveillance systems require accurate monitoring of objects if/when they violate perimeter restrictions. In this paper, we seek a solution for depth imaging of such objects of interest at high accuracy (few tens of cm) over extended ranges (up to 300 meters) from a *single vantage point*, such as a pole mounted platform. Unfortunately, the rich literature in depth imaging using camera, lidar and radar in isolation struggles to meet these tight requirements in real-world conditions. This paper proposes Metamoran, a solution that explores long-range depth imaging of objects of interest by fusing the strengths of two complementary technologies: mmWave radar and camera. Unlike cameras, mmWave radars offer excellent cm-scale depth resolution even at very long ranges. However, their angular resolution is at least $10\times$ worse than camera systems. Fusing these two modalities is natural, but in scenes with high clutter and at long ranges, radar reflections are weak and experience spurious artifacts. Metamoran's core contribution is to leverage image segmentation and monocular depth estimation on camera images to help declutter radar and discover true object reflections. We perform a detailed evaluation of Metamoran's depth imaging capabilities in 400 diverse scenarios. Our evaluation shows that Metamoran estimates the depth of static objects up to 90 m away and moving objects up to 305 m away and with a median error of 28 cm, an improvement of 13$\times$ over a naive radar+camera baseline and 23$\times$ compared to monocular depth estimation.**

I. Introduction

Surveillance and geo-fencing are classic problems in robotics, where one seeks to identify specific objects of interest and observe if they violate perimeter restrictions. Moving beyond short range applications where depth cameras thrive [1], we ask the question, "what does it take to build a single fixed vantage point sensing solution that can create accurate depth images of objects at long ranges?" A single vantage point solution allows for quick deployment in scenarios where infrastructure is hard to come by, with minimal calibration. For example, one can imagine a single pole-mounted platform that monitors people or vehicles trespassing large private areas or drones entering no-fly zones.

Several sensors such as cameras [2], [3], [4], lidars [5] and radars [6] have been used for single vantage point depth imaging. Monocular solutions [2] experience tens of meters of error for objects beyond $\sim$30 m. Standalone lidar solutions provide sparse but accurate depth estimates over ranges of 100-200 m [7]. Indeed, sensor fusion approaches combining camera and lidar [8], [9], [10] have recently been proposed to generate high angular resolution, accurate depth images. However, real world lidar data fails to detect certain objects between 30-50 m depending on object reflectivity characteristics, orientation and ambient sunlight (Sec. VI). This motivates us to explore other sensing modalities such as radar which has become popular owing to availability of large frequency spectrum in millimeter wave (mmWave) frequencies (60 and 77-81 GHz).

One of the most appealing features of mmWave radars is its high bandwidth, which enables object detection often as far as 150-300 m at cm-scale depth resolutions. Yet, mmWave radars, by themselves, are not a high angular resolution depth imaging solution because of the limited number of antennas that are packed on a small form factor radar. The best commercially-available radars achieve an angular resolution of 1.5° [6], which is at least $10\times$ worse than cameras. Previous works have compensated for the poor angular resolution by fusing with camera [11], [12], [13] but only for short range (10-20 m) — where their impressive operating range and depth resolutions are not fully utilized.

This paper considers the unique problem of mmWave radar and camera sensing for long-range depth imaging of specific objects of interest (both static and dynamic). This is challenging because unlike systems that operate over short ranges, the first peak detected in radar doesn't necessarily correspond to the detected object in the image. This is primarily because of overwhelming reflections from ambient but out of interest objects that can clutter the scene. Static cluttered scenes are more challenging because traditional Doppler processing doesn't help. We propose Metamoran which leverages semantic information from a monocular camera to help the radar disambiguate between objects of interest and clutter.

An intuitive starting point for Metamoran to eliminate unnecessary clutter is to segment the camera image (see Fig. 1a) and use the radar to look for peaks *only* within the angular span $(\theta_1, \theta_2, \phi_1, \phi_2)$ occupied by the objects of interest as identified in the segmentation output. This helps the radar ignore reflections from out of interest objects. A practical challenge in designing this arises because of strong clutter from objects such as buildings, lamp posts and fences. The presence of a strong reflector creates undesirable side lobes that spread across the angular axis (see Fig. 1c).

A. Prabhakara, C. Li, A. C. Sankaranarayanan, A. Rowe and S. Kumar are with Carnegie Mellon University, Pittsburgh PA 15213 USA. Email: {aprabhak@andrew., chaoli2@andrew., saswin@andrew., agr@ece., swarun@} cmu.edu

D. Zhang is with Applied Physics Laboratory Johns Hopkins University, Laurel MD 20723 USA. Email: diana.zhang@jhuapl.edu

S. Munir is with Bosch Research and Technology Center, Pittsburgh PA 15222 USA. Email: sirajum.munir@us.bosch.com

Corresponding Author: Akarsh Prabhakara.

Akarsh Prabhakara and Diana Zhang are co-first authors.

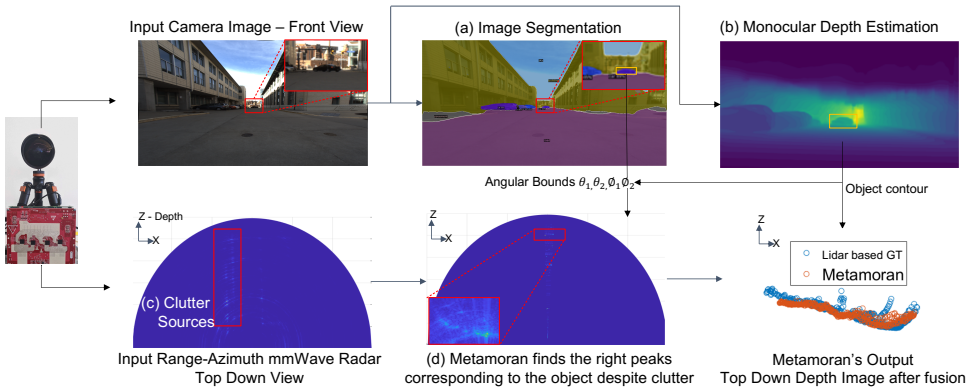

Fig. 1. Metamoran takes camera and mmWave radar signals as input, tackles clutter and produces high resolution depth images of objects of interest by co-optimizing radar processing with inputs from computer vision techniques such as image segmentation and monocular depth estimation.

This implies that even after segmenting radar to only angles where the object of interest is present, side lobes from strong reflectors at other angles create ghost peaks within the angles of interest. This is extremely critical as depth estimation can have significant errors if ghost peaks are selected.

To tackle this challenge, we design a new radar processing algorithm which selects the correct peaks by leveraging the camera image. Our key idea is unlike clutter, the peaks in radar images of an object of interest *look like* the object. For example, the contour of a car is quite evident in Fig. 1d. Using image segmentation and monocular depth estimation, we obtain these contours which capture object type, angles of interest, and objects' internal depth variation, although not on an absolute scale. We use this estimated contour to synthesize templates of radar signal that would have been measured if no other objects were in view. Our processing uses this template to suppress clutter, maximize the signal strength of objects of interest and find the true peak in radar. The depth so obtained is cm precise owing to the high depth resolution of mmWave radars. The rest of our pipeline is designed to further eliminate the side lobes and use the estimated depth to create a high angular resolution, accurate depth image.

Contributions: We make the following contributions:

- A radar processing system that combines semantic camera data to find objects of interest in mmWave signals, in high clutter and resulting undesirable radar side-lobes (Sec. III).
- A pipeline that fuses high angular resolution monocular depth estimation with accurate radar depth estimation to create a depth image (Sec. IV).
- A detailed implementation including extensive raw radar data[1] (static and dynamic scenes totaling 125 GB) along with high resolution, raw camera images and ground truth based on lidar in diverse scenes outdoors (Sec. V).
- Evaluation of Metamoran in various high-clutter environments to demonstrate substantial improvements in long range depth imaging. (Sec. VII).

While this paper is focused on robotic surveillance as the key application use case, radar-camera fusion for depth imaging is valuable for other wide ranging applications: autonomous driving and enhanced robotic perception. This study provides the tools needed to enable a richer exploration of robotic use cases for hybrid mmWave and camera sensing.

[1]https://www.witechlab.com/metamoran.html

II. Related Work

mmWave Radar Imaging: With the proliferation of mmWave radar devices, radio frequency imaging, which used to be prevalent in lower frequencies [14], is reaping benefits from the wide bandwidth available at mmWave frequencies. More recently, radar angular resolution is being improved using deep learning [15], [16] and through synthetic aperture [17], [18], [19] for a variety of contexts including high fidelity through wall/obstruction imaging. While complementary, these solutions are not designed to produce high-resolution depth images at extended long distances.

Radar-Camera Fusion: Camera and radar fusion has been proposed for robust object perception and detection [12], target tracking [11], [13], obstacle detection [20], [21], [22] and autonomous driving [23], [24]. While the problem of depth imaging is different, it is important to note that some of the older works use mechanical scanning radar or electronic scanning with a very narrow FoV which leads to denser and less cluttered output. More recently with the availability of point cloud radar data through nuScenes, deep learning fusion techniques for 3D object detection [25], [26] have been proposed. In contrast to the point cloud data, we collect raw radar signals because our processing algorithms are not learning based and they rely on using the phase and amplitude of the time series signal, and not just point cloud intensity. Few other works also create their own dataset [27], but they operate in short ranges of 0-25 m. Beyond drawing bounding boxes for 3D object detection, in this work our problem definition involves *imaging*, that is obtaining the depth variation across RGB pixels for an object of interest.

Radar Clutter Suppression: Traditional algorithms to tackle radar clutter include Doppler processing to detect moving objects in a static background scene. Here, we are interested in both static and dynamic objects. While dealing with static clutter when detecting a static object, the clutter profile must be first computed. Some of these techniques include subspace projection [28] and adaptive filtering [29]. These techniques only use radar information. In this work,

by building a radar and camera fusion system, we not only leverage radar data but also camera semantic data to search for object of interest and suppress clutter.

III. Accurate Depth Estimation

The first step in performing Metamoran's depth estimation isolates the object of interest in the radar image in high clutter environments using information from the corresponding camera image. This step is crucial in removing the impact of clutter in the radar image that may otherwise be misled by non-existent or irrelevant objects in the scene. The specific approach we use for camera image pre-processing is panoptic segmentation, which identifies spatial bounds and attaches semantically meaningful labels to objects in the image.

Image Segmentation Pre-processing: We perform image segmentation using pretrained Detectron2 [30]. This model has been previously trained on several objects including cars and persons in various short and long range environments. We use these types of objects as our primary test subjects without additional model tuning. The output of image segmentation is a segmentation mask (the angular bounds of an object), a semantic label for the mask (e.g. car, person, etc.) and instance ID (to identify specific cars, etc.).

Radar Processing Pipeline: We then use the output of image segmentation to carve out objects of interest in the radar image (i.e. heatmap as in Fig 2). For example, if a car lies in line of sight between $-5°$ and $0°$, the radar heatmap seen in Fig 2, is truncated to these angular limits. Assuming that the object is in line of sight with respect to radar, in an ideal world, within the angular limits there should only exist peaks corresponding to the object of interest and nothing else. However, in high clutter environments, strong, out of interest reflectors which can even lie outside the angular limits, tend to leak their signal into angles of interest (see Fig. 2). Such strong reflectors like buildings, tend to spread out their signal along the azimuth axis in a sinc-like fashion with decreasing side lobe levels. These side lobes affect across all angles at the same range bin and show up as false peaks within the angle of interest. A naive radar camera fusion would end up choosing these false peaks. The rest of this section describes our approach in accurately detecting peaks even in the presence of high clutter.

A. Computing Object Depth

After segmenting the radar image to the desired angles of interest, Metamoran's key next step is a novel radar processing algorithm, which searches for peaks that resemble the shape of the object. Our idea is to first build an approximate shape of the object by leveraging camera data and then look for this shape in the radar image. Specifically, we use monocular depth estimation, a classic image processing solution that captures the relative depth variation (i.e. shape) of all objects in the entire scene in an RGB-D depth image. Our specific choice of monocular depth estimation on camera data is AdaBins [2] (see example output in Fig. 3) which is trained on extensively used KITTI dataset [31]. While it is highly accurate in capturing the depth variation of an object, a downside of monocular depth imaging is that it is poor in terms of absolute depth accuracy, when compared to cm-scale accuracy of radars. We therefore seek to capture the absolute depth information of the object, whose shape we found through monocular depth estimation, using a radar processing pipeline that we describe below.

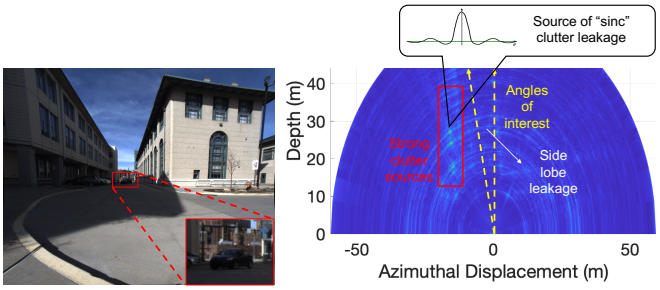

Fig. 2. Metamoran tackles the overwhelming clutter from strong reflectors such as buildings which dwarf the reflection from desired objects.

Mathematically, let P be the binary mask which corresponds to the pixels of the object of interest as obtained through image segmentation. The monocular depth estimate obtained from [2] can be captured in a matrix M, which is basically the "D" slice from RGB-D image. We can then obtain the approximate 3D shape S of the object by element wise multiplication of M and P. S is now a matrix which is largely 0, but in pixels where object is present, it has monocular depth estimates. Because we are using a 2D radar (range and azimuth) with a narrow elevation FoV, rather than using the full 3D shape, we extract a contour C as essentially a row chosen from S. The row index translates to elevation angle and if radar and camera are co-located, we simply choose the centermost row. The column index translates to azimuth angle and we convert the column indices to appropriate azimuth angles. Choosing non-zero elements in this row, we have a point cloud that can be indexed by azimuth angle and depth value. The contour captures 2D shape of the object – that is, depth variation over azimuth (see Fig. 3). We can then transform these coordinates to $C(x, z)$. The depth values obtained so far are not accurate, because they are still monocular estimates.

With the obtained shape, Metamoran next models the reflections that the radar would have received if only points in

Algorithm 1: Depth Estimation Algorithm

Input : Image Segmentation Object Mask, P
Monocular Depth Estimation, M
Raw I/Q Radar capture, h

$S = M \cdot P$ // Approximate 3D shape of object
$C(x, z) =$ GetShapeContour$(S(x, y, z))$
for *depth* d **do**
  $h^d_{template} =$ShiftToDepth$(C(x, z), d)$
  $P(d) = corr(h^d_{template}, h)$ // Matched Filter
$d^* = \underset{d}{\text{argmax}} P(d)$ // Depth Estimate
Output: d^* // Depth Estimate

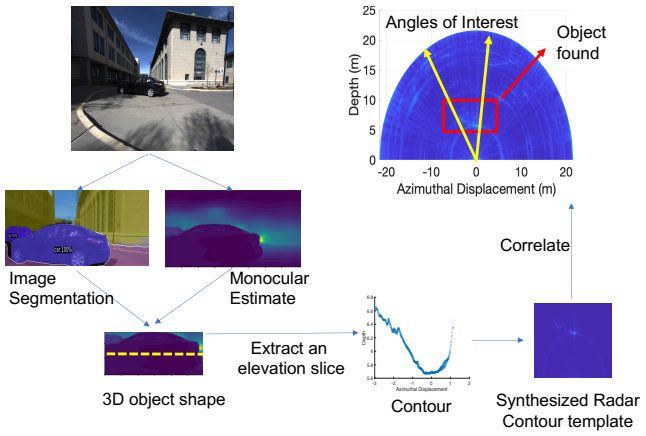

Fig. 3. Metamoran uses information from camera image segmentation and monocular depth estimation to obtain a coarse contour of the object of interest. It then uses this contour to perform correlation to find the object in radar image amidst clutter and thereby estimate its depth accurately.

C were present without any clutter. Metamoran synthesizes the contour signal template by modeling each point on the contour $C(x, z)$ as a point reflector. In its simplest form, one can obtain this point's contribution to the synthesized Frequency Modulated Continuous Wave signal as [32]:

$$h_{template;i}(n) = \alpha e^{j\frac{4\pi D_i}{\lambda}} e^{j2\pi\frac{D_i}{D_{max}}n} \quad (1)$$

where, α is the amplitude of the received signal, D_i is the distance between (x_i, z_i) and the radar antenna, D_{max} is the maximum distance that the radar is configured to operate, λ wavelength, n indexes digital time series samples, and $j = \sqrt{-1}$. Superposing each point's contribution we obtain the overall signal template for the entire contour as $h_{template}$.

Finally, Metamoran explores at what absolute depth the shape-contour in C is present within the radar image. To compute this depth, Metamoran shifts the point cloud differently such that absolute depth of the closest point is at different d, synthesizing a new template each time $h^d_{template}$, and applies a matched filter to obtain $P(d)$ – the correlation of the contour template, at each possible depth d, with respect to the measured radar signal. Mathematically, if h is the original received radar signal, we have:

$$P(d) = corr(h^d_{template}, h)$$

This correlation is performed across h measured at each radar antenna and then aggregated. We then report the depth estimate of this object as the value of d that corresponds to the maximum of $|P(d)|$, i.e.

$$d^* = \arg\max_d |P(d)|$$

With d^*, we know accurately the closest depth of the object with respect to the radar. The matched filtering operation essentially searches for the objects' shape and promotes the signal strength corresponding to the objects' reflections but not clutter. The computational complexity of such a method with D different depths of interest and N length vector h would be $\mathcal{O}(DN^2)$. Additionally this complexity can be reduced by converting correlations to FFTs and searching over different depths hierarchically from coarse to fine.

In this subsection, we showed that by using segmentation and monocular depth estimation, we can pick the objects' peaks accurately. To further help finding objects' peaks in cluttered conditions, the following subsection describes how camera information can also be used to suppress the clutter.

B. Clutter Suppression

Clutter due to strong reflections from undesired objects can impede Metamoran. For instance, even if an undesired object is at an azimuth significantly different from the desired object, it's side lobes can create ghost peaks that causes interference. Worse still, some reflectors may be orders of magnitude stronger than our desired object, and thus even their side lobes can dwarf our objects of interest. Fig. 2 shows an example of a highly cluttered scene. Our objective here is to remove unwanted clutter to focus on the object of interest. While the shape-correlator based detector was designed to avoid choosing ghost peaks, if the object of interest is dwarfed by very strong reflections, then these can trigger the correlator detector and result in a faulty depth estimates. Therefore, one must perform a declutter phase prior to applying Metamoran's shape correlator.

Specifically, in Metamoran, we look for semantic objects that are usually strong reflectors such as buildings, fences and lamp posts using the camera segmentation output. For each such strong peak outside the angle of interest, we treat it as a point reflector at a certain detected range and azimuth, and synthesize a template following Eqn. 1, which captures its contribution to the measured signal. A key point to note is that because these are strong reflectors, α of the template is chosen to be equal to the peak value. This template is then subtracted from the measured radar signal. We iterate over several such peaks many times until the magnitude of the peaks in the angles of interest are comparable to the expected magnitude of an object reflection. This is analogous to successive interference cancelation in RF communication [33]. What this process accomplishes is the removal of side lobes from these large peaks within our angles of interest – thereby enabling robust object peak detection. For P peaks this algorithm is essentially subtracting the template P times giving a computational complexity of $\mathcal{O}(P)$.

IV. Depth Imaging

We note that our current description of Metamoran's algorithm provides only one depth value per object template, i.e. one depth per object. In practice, we deal with extended objects and we would require a depth image across the object. We could use local peaks from the clutter-suppressed radar image near the peak depth value obtained from shape-correlation algorithm. But, the point cloud so obtained is very sparse and only becomes sparser with increasing object distances. To mitigate this, we rely on fusing the sparse but accurate radar peaks obtained from shape-correlator with the output of the dense camera-image based monocular depth estimation discussed previously (AdaBins [2], see Fig. 3). However, two problems persist in realizing this fusion.

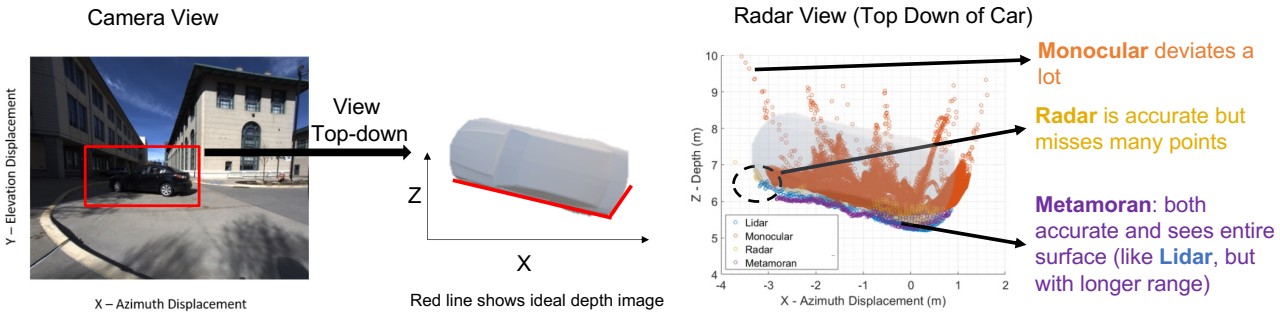

Fig. 4. **Metamoran vs. Radar and Monocular Estimation:** A qualitative comparison of the depth images in a clutter free, close range scene shows standard radar to be coarse in azimuth resolution, monocular to have significant absolute depth offsets but great azimuth diversity, and Metamoran which leverages rich shape information from image pre-processing to generate an accurate, dense depth image. At longer ranges, monocular depth estimates would deviate several meters and radar would get sparser.

Correcting Absolute Errors: While monocular depth estimation may often correctly return the *relative* depths between parts of a large object such as a car, it often makes large errors in *absolute* depths, particularly for objects at long ranges (see Table I). To fuse monocular and shape-correlator output, we want them to be absolute depth aligned. To resolve this, we rely on more accurate absolute depth estimate from radar obtained in Sec. III-A and shift monocular depth point cloud such that the closest point is at d^*.

Correcting Relative Errors: After aligning the monocular depth estimates with the sparse point cloud from Metamoran's shape-correlator, a naive way to fuse this would be consider all points from both modalities. But, as seen in Fig. 4, edges of monocular estimates tend to deviate quite significantly from the primary contour outline of the object. This could be because of imprecise segmentation or that monocular depth estimation often struggles with objects that do not have significant variation in color with respect to the background or sharp edges that intuitively simplifies depth estimation [34], [35]. If fused as is, one would experience higher errors as expected from monocular depth estimation. It is therefore important to select points from the aligned monocular depth estimates that only lie along the primary contour outline and reject outliers. We note that the number of points detected per azimuth bin in monocular estimates fall off sharply at the edges where our outliers of interest lie. By using a simple threshold based outlier detection, we identify points which actually lie along the primary contour. Upon fusing selected monocular depth estimate points and sparse point cloud from Sec. III-A, we obtain a depth image, that resembles ground truth lidar and outperforms different algorithms using either of the two modalities in terms of azimuth resolution and depth accuracy (see Fig. 4).

V. Implementation and Evaluation

System Hardware: Metamoran is implemented using a FLIR Blackfly S 24.5MP color camera and a TI MMWCAS-RF-EVM RADAR (see Fig. 5). We operate the radar at 77-81 GHz in a TDM-MIMO mode. This radar has a theoretical range resolution of 3.75-60 cm, depending on max range. The radar also has 86 virtual antennas spaced out along the azimuth axis with a narrow elevation FoV, which provides a theoretical azimuth resolution of $1.4°$. This is at least an order of magnitude worse than cameras and lidars. Unlike fusion approaches which rely on processed point clouds [36], this radar supports logging raw complex samples which is critical for our processing. The whole hardware system is kept about 1 m above ground level during data collection.

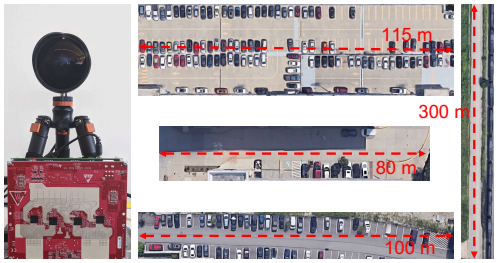

Fig. 5. **Metamoran's Hardware Platform:** We use a FLIR Blackfly S 24.5MP color camera and a TI MMWCAS-RF-EVM mmWave radar. We deploy our system in outdoor spaces with high clutter.

Testbed and Data Collection: We test this system in a variety of 400 outdoor scenes such as parking lots and roads at distances ranging up to 320 m from objects of interest. These environments have rich clutter sources arising due to buildings, street lamps, fences, trees, trains, out of interest parked cars and pedestrians. Fig. 5 shows four candidate locations in the area surrounding a university campus.

Ground Truth: We collect ground truth data using a Velodyne Puck Lidar (VLP-16), which generates 3D point clouds, with fine angular resolution and 3 cm ranging error. While this lidar is rated for up to 100 m, in practice, on a sunny day, we found the Puck collected data with sufficient point cloud density only until about 20-30 m. Therefore, for ranges beyond 20 m, we surveyed a point closer to the object of interest and placed the lidar at that point.

Baselines: We compare Metamoran with two baselines that use the same hardware platforms: (1) *Monocular Depth Estimation:* We use state-of-the-art monocular depth estimation algorithm [2]. (2) *Naive fusion of Camera and Radar:* We use image segmentation to obtain the azimuth spanned by object of interest. We perform standard radar processing for FMCW radar, and bound the output to the azimuth span and then pick the strongest reflector as the object.

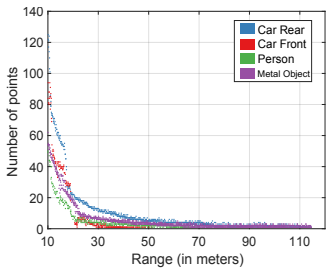

Fig. 6. Lidar stops detecting non retro-reflective objects sooner than rated.

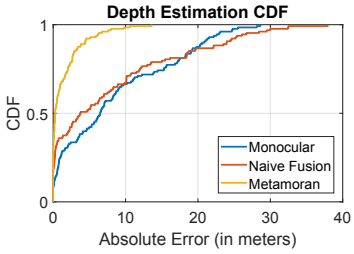

Fig. 7. CDF of depth errors shows our depth estimation at long ranges in clutter.

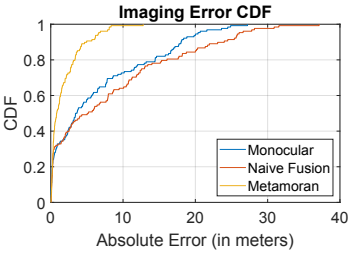

Fig. 8. CDF of depth imaging errors shows our performance despite clutter.

Method	All classes		Across object types						Across range bins					
			Car		Metal Objects		Person		0-20m		20-40m		40-60m	
	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD
Monocular [2]	6.50	7.97	6.06	7.12	9.56	8.11	5.16	9.23	0.60	1.30	6.75	3.41	19.43	4.00
Naive Fusion	3.75	9.34	2.07	8.91	7.25	8.37	5.04	10.62	0.18	2.19	7.00	7.19	15.95	11.09
Metamoran (ours)	**0.28**	**2.35**	**0.02**	**2.28**	**0.85**	**1.89**	**0.57**	**2.75**	**0.06**	**0.94**	**0.71**	**3.05**	**1.27**	**1.94**

TABLE I

DEPTH ESTIMATION ERRORS: METAMORAN OUTPERFORMS BOTH BASELINES FOR A VARIETY OF OBJECTS WITH DIFFERENT REFLECTIVITIES, IN DIFFERENT ORIENTATIONS AND LONG RANGES. MAE- MEDIAN ABSOLUTE ERROR (IN METERS) STD- STANDARD DEVIATION (IN METERS).

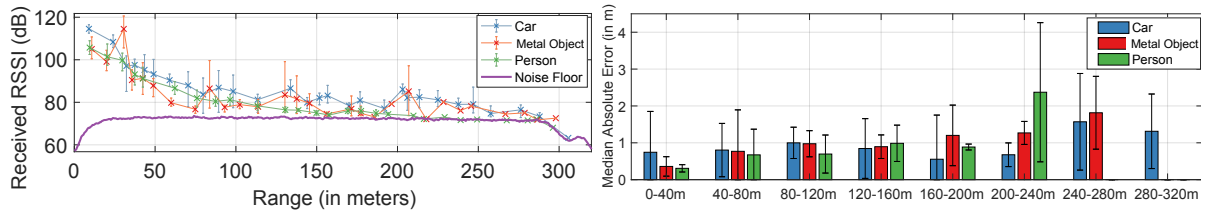

Fig. 9. Metamoran leverages radar doppler processing to detect and range moving objects even at 300 m where their RSSI (Received Signal Strength Indicator) is about 20dB lower than the surrounding static clutter.

Objects of interest Selection: We select a car, a person and metal objects (such as stop signs) for use as our objects of interest as these are useful for varied applications, including surveillance. Our choice also provides a variety of reflectors in size, shape, and reflectivity. We note that in Metamoran, we detect both static and moving objects. Indeed, static objects are much more challenging to detect in radar because Doppler filtering cannot be used to remove clutter.

Calibration: We note that data Metamoran collects requires both internal calibration of the components as well as external calibration between the camera and radar. Internally, our mmWave radar is calibrated using a corner reflector placed at 5 m [37]. The camera intrinsics are measured by taking many photos of a checkerboard to remove fisheye distortion. Externally, even though both sensors are co-located, they are at a small relative vertical displacement of 15 cm and relative rotations. Prior to fusing, these offsets are compensated to ensure consistency.

VI. MICROBENCHMARKS

Method: We study lidar's maximum detection range and the distance at which we have sufficient point cloud density for use as ground truth. We collect the lidar data of objects at different ranges but the same orientation towards the lidar.

Results: We noticed that the maximum detection range depends on object reflectivity characteristics. We see in Fig. 6 that only one point from the front of the car (without a license plate) is detected between 30-50 m. Depending on the color of the paint and orientation of the car, we observed that the front would stop being detected even between 25-30 m. This could largely be because of mirror-like reflectivity causing reflections to never return to lidar. However, objects with retroreflective surfaces such as rear of the car with license plate and our chosen metallic objects are detectable up to 114.3 m and 114.6 m respectively. A person being a diffuse scatterer is detectable up to 64 m. Surveillance applications cannot afford to make any assumptions on the mirror like/retroreflectivity of objects. Although, specular properties have been investigated in radar [38], we show that without making any reflectivity assumption Metamoran can detect all orientations of car, metallic objects and person effectively. Note that point cloud density drops drastically as objects move away. We pick 20 m as the range with sufficient point cloud density. For collecting ground truth beyond 20 m, we move the lidar closer to the object.

VII. RESULTS

A. Depth Estimation

Method: We first evaluate depth estimation accuracy by collecting data samples in varying lighting conditions at 4 clutter rich sites. Static objects were positioned from 3-90 m and were placed in various orientations with respect to radar and camera setup. Data was collected in 3 range bins at

Method	All classes		Across object types						Across range bins					
			Car		Metal Objects		Person		0-20m		20-40m		40-60m	
	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD
Monocular [2]	3.53	7.20	3.16	6.05	7.58	7.56	4.15	8.46	**0.17**	**0.83**	3.83	2.90	17.08	4.48
Naive Fusion	5.13	9.11	2.27	8.39	7.77	8.38	7.39	10.50	0.30	1.93	6.08	7.70	14.58	9.28
Metamoran (ours)	**0.82**	**2.26**	**0.70**	**2.39**	**1.17**	**1.77**	**1.07**	**2.42**	0.25	0.90	**1.50**	**2.72**	**1.77**	**2.00**

TABLE II

DEPTH IMAGING ERORS: METAMORAN GENERATES HIGH RESOLUTION, ACCURATE DEPTH IMAGES OF OBJECTS AT LONG RANGES. MAE- MEDIAN ABSOLUTE HAUSDORFF DISTANCE (IN METERS) STD- STANDARD DEVIATION (IN METERS)

different resolutions: 4.2 cm at 0-20 m, 11.6 cm at 20-60 m, 21 cm at 60-90 m. The primary bottleneck in maintaining 4 cm range resolution at long ranges is the TDA2SX SoC capture card on the MMWCAS board – it can handle at most a data width of 4096, corresponding to 512 complex samples per receiver. Thus at longer ranges, we can't utilize the full potential of mmWave range resolution.

Depth error is measured at the object point which is closest to the radar. For each baseline and Metamoran, we compare median absolute error (MAE) with respect to lidar. Below, we represent three sets of depth errors: (1) across object categories (2) across different range bins (3) overall error distribution. Across all experiments, we find that Metamoran significantly outperforms the baselines.

Object Results: Table. I shows the median error in depth across objects of interest. We see lowest error for car across the board due to a combination of factors: car is our strongest reflector, offers multiple points on its surface to reflect radar signals due to its size and thereby a high radar cross section. We see performance further degrade with the weaker reflectors. Metallic objects have a higher error compared to person because although it's more reflective to radar, it suffers from specular reflections.

Range Results: Table. I also shows the median error in depth across range bins. As expected, accuracy across all approaches deteriorates with range due to weaker received signals. Here, we can clearly see monocular estimation suffer beyond the 20-40 m bin and also see the effects of clutter rendering naive fusion erroneous. Metamoran which tackles clutter continues to do well even in 40-60 m bin. For experiments performed in the 60-90 m range bin, our baselines encounter huge errors. Because of extremely low received power, metallic objects and person are no longer detectable even with the assistance of Metamoran. However, Metamoran detects cars up to 90 m with a MAE of 1.1 m and standard deviation of 2.45 m.

CDF Results: Fig. 7 shows the overall distribution of our depth errors. Metamoran has a median error of **0.28 m** across all collected data. Metamoran clearly outperforms the baselines to accurately detect a variety of objects in different orientations and in high clutter environments. Between monocular and naive fusion, we can see that naive fusion benefits from the fusion and has a lower MAE but suffers due to high clutter and has long tailed distribution.

B. Extremely Long Ranges

Method: To evaluate the maximum detection range of Metamoran we perform the following experiments. We have already found the limit for static objects in high clutter as 90 m. We now leverage Doppler processing that Frequency Modulated Continuous Wave (FMCW) radars are capable of to detect moving objects. Although FMCW is very popular for today's radars, modern lidars are largely time-of-flight based. Therefore, in building a fusion system, radar Doppler processing brings unique advantages. To evaluate the true radar detection ability we collect data for moving car, person and metallic objects up to 320 m. For every data snapshot collected, the object moves at a slow speed towards the radar. For these ranges, we collect the data at 30 cm range resolution up to 120 m and 60 cm resolution up to 320 m.

Results: We see in Fig. 9 that the received signal strength drops consistently until 305 m, when it hits the noise floor. The signal strength variations are particularly large for metallic objects because they are sensitive to orientation with respect to radar. The person is detected up to 229 m, metallic objects up to 298 m and car up to 305 m. Although the reflection from these objects at long ranges are extremely small compared to background clutter, just because they are moving, Doppler processing can still detect the objects. We also see that in Fig. 9 the depth errors increase with distance as expected. Because of the radar resolution of 60 cm, at these long ranges, even for cars the errors can reach 1.5 m. Given enough signal integration, these errors should decrease and reach the resolution limit. As long as the radar detects these objects and estimates depth accurately, Metamoran's depth imaging algorithms are still applicable albeit with the help of a pan, tilt, zoom camera to get a high resolution camera image of the object for fusion.

C. Depth Imaging

Method: To compute high resolution depth images, we implement the method in Sec. IV. In contrast to Sec. VII-A which only computed depth errors, here we want to characterize accuracy for a point cloud obtained from the baselines monocular depth estimation and naive fusion of camera and radar, and our system against lidar point clouds. Data was collected similar to Sec. VII-A.

To compare two point clouds A and B, we use a modified version of Hausdorff distance [39] as follows:

$$\min\left\{\underset{a\in A}{\text{median}}\{\underset{b\in B}{\min}\{d(a,b)\}\}, \underset{b\in B}{\text{median}}\{\underset{a\in A}{\min}\{d(b,a)\}\}\right\}$$

where $d(a, b)$ is the distance between points a and b. Hausdorff distance is popularly used in obtaining similarity scores between point clouds. Intuitively, this metric measures the median distance between any two points in the point cloud. The lower the distance, the more similar the point clouds.

Results: Table. II shows our depth imaging results. Trends in imaging results largely follow those in depth estimation, as problems with detection propagate through the pipeline. Metamoran outperforms both baselines across all categories, except 0-20 m where all three methods produce comparable results. Fig. 8 shows CDF of errors in depth imaging. Metamoran has a median absolute error of **0.82 m** across all collected data. We note that monocular depth estimation outperforms naive fusion unlike in Sec. VII-A. This once again shows that, while monocular benefits from large azimuth span of points for extended objects like cars, high clutter makes naive fusion pick wrong depth estimates which lead to larger imaging errors. We also find that for experiments performed in the 60-90 m range bin, Metamoran successfully images static cars with a MAE of 1.98 m and standard deviation of 1.7 m.

VIII. Conclusion

This paper develops Metamoran, a mmWave radar and camera based system that achieves high resolution depth images for objects at long ranges and in high clutter environments. Metamoran's secret sauce is in leveraging processed camera information to declutter the scene, eliminate false peaks and identify the right peaks. Metamoran also uses the detected peak and processed camera information to create a high resolution depth image of desired objects. Metamoran was evaluated extensively up to 300 m. The resulting dataset is extremely valuable to the robotics community as it offers ground truth lidar, camera and raw radar data. We believe there is a strong role for Metamoran's radar-camera fusion, as a complementary approach to lidar, in deploying rich robotic applications such as robotic and autonomous vehicular navigation and sensing, while ensuring overall resilience to occlusions and weather conditions.

Acknowledgements: This research was supported by the National Science Foundation (1823235, 2106921, 2030154 and 2007786), Bosch and DARPA TRIAD.

References

[1] StereoLabs, https://www.stereolabs.com/zed-2/, 2020.
[2] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *CVPR*, 2021.
[3] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.
[4] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv:1812.11941*, 2018.
[5] Velodyne, https://velodynelidar.com/products/puck/, 2021.
[6] Texas-Instruments, https://www.ti.com/tool/MMWCAS-RF-EVM.
[7] N. Schneider, L. Schneider, P. Pinggera, U. Franke, M. Pollefeys, and C. Stiller, "Semantically guided depth upsampling," in *German conference on pattern recognition*, 2016.
[8] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *ICRA*, 2019.
[9] S. S. Shivakumar, T. Nguyen, I. D. Miller, S. W. Chen, V. Kumar, and C. J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," in *ITSC*, 2019.
[10] L. Chen, Y. He, J. Chen, Q. Li, and Q. Zou, "Transforming a 3-d lidar point cloud into a 2-d dense depth map through a parameter self-adaptive framework," in *IEEE Trans. on ITS*, 2016.
[11] R. Zhang and S. Cao, "Extending reliability of mmwave radar tracking and detection via fusion with camera," in *IEEE Access*, 2019.
[12] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji, and G. Xing, "millieye: A lightweight mmwave radar and camera fusion system for robust object detection," in *IoTDI*, 2021.
[13] A. Sengupta, F. Jin, and S. Cao, "A dnn-lstm based target tracking approach using mmwave radar and camera sensor fusion," in *IEEE NAECON*, 2019.
[14] D. Huang, R. Nandakumar, and S. Gollakota, "Feasibility and limits of wi-fi imaging," in *ACM Sensys*, 2014.
[15] J. Guan, S. Madani, S. Jog, S. Gupta, and H. Hassanieh, "Through fog high-resolution imaging using millimeter wave radar," in *CVPR*.
[16] S. Fang and S. Nirjon, "Superrf: Enhanced 3d rf representation using stationary low-cost mmwave radar," in *EWSN*, 2020.
[17] A. Prabhakara, V. Singh, S. Kumar, and A. Rowe, "Osprey: A mmwave approach to tire wear sensing," in *ACM MobiSys*, 2020.
[18] M. E. Yanik and M. Torlak, "Near-field mimo-sar millimeter-wave imaging with sparsely sampled aperture data," in *IEEE Access*, 2019.
[19] C. M. Watts, P. Lancaster, A. Pedross-Engel, J. R. Smith, and M. S. Reynolds, "2d and 3d millimeter-wave synthetic aperture radar imaging on a pr2 platform," in *IROS*, 2016.
[20] V. John, M. Nithilan, S. Mita, H. Tehrani, R. Sudheesh, and P. Lalu, "So-net: Joint semantic segmentation and obstacle detection using deep fusion of monocular camera and radar," in *PSIVT*, 2019.
[21] S. Han, X. Wang, L. Xu, H. Sun, and N. Zheng, "Frontal object perception for intelligent vehicles based on radar and camera fusion," in *Chinese Control Conference*, 2016.
[22] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," in *Sensors*, 2020.
[23] H. Cho, Y. Seo, B. V. K. V. Kumar, and R. R. Rajkumar, "A multi-sensor fusion system for moving object detection and tracking in urban driving environments," in *ICRA*, 2014.
[24] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," in *Trans. on ITS*, 2007.
[25] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *WACV*, 2021.
[26] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *CVPR*, 2021.
[27] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization," in *IEEE JSTSP*, 2021.
[28] J. E. Palmer and S. J. Searle, "Evaluation of adaptive filter algorithms for clutter cancellation in passive bistatic radar," in *IEEE Radar Conference*, 2012.
[29] R. Solimene, A. Cuccaro, A. Dell'Aversano, I. Catapano, and F. Soldovieri, "Ground clutter removal in gpr surveys," in *IEEE Journal of STAEO and Remote Sensing*, 2013.
[30] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.
[31] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.
[32] Texas-Instruments, "The fundamentals of millimeter wave radar sensors," https://www.ti.com/lit/wp/spyy005a/spyy005a.pdf, 2021.
[33] P. Patel and J. Holtzman, "Analysis of a simple successive interference cancellation scheme in a ds/cdma system," in *IEEE JSAC*, 1994.
[34] A. Saxena, S. H. Chung, A. Y. Ng, *et al.*, "Learning depth from single monocular images," in *NIPS*, 2005.
[35] M. A. Reza, J. Kosecka, and P. David, "Farsight: Long-range depth estimation from outdoor images," in *IROS*, 2018.
[36] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *IEEE SDF*, 2019.
[37] "User's guide: Ti mmwave studio cascade," 2018.
[38] K. Bansal, K. Rungta, S. Zhu, and D. Bharadia, "Pointillism: Accurate 3d bounding box estimation with multi-radars," in *ACM Sensys*, 2020.
[39] pdal.io, https://pdal.io/apps/hausdorff.html.