Learned Depth Estimation of 3D Imaging Radar for Indoor Mapping

Ruoyang Xu, Wei Dong, Akash Sharma, and Michael Kaess

***Abstract*— 3D imaging radar offers robust perception capability through visually demanding environments due to the unique penetrative and reflective properties of millimeter waves (mmWave). Current approaches for 3D perception with imaging radar require knowledge of environment geometry, accumulation of data from multiple frames for perception, or access to between-frame motion. Imaging radar presents an additional difficulty due to the complexity of its data representation. To address these issues, and make imaging radar easier to use for downstream robotics tasks, we propose a learning-based method that regresses radar measurements into *cylindrical* depth maps using LiDAR supervision. Due to the limitation of the regression formulation, directions where the radar beam could not reach will still generate a valid depth. To address this issue, our method additionally learns a 3D filter to remove those pixels. Experiments show that our system generates visually accurate depth estimation. Furthermore, we confirm the overall ability to generalize in the indoor scene using the estimated depth for probabilistic occupancy mapping with ground truth trajectory. The code and model will be released[1].**

I. Introduction

One of the reasons for recent successes in simultaneous localization and mapping (SLAM) systems is a thorough understanding of the sensors used for sensor fusion. Consequently, modern state estimation systems rely on a *standard suite* of sensors that includes visual, inertial and laser based sensors, sparing a few exceptions such as the Doppler velocity log (DVL), and thermal cameras in specific applications such as underwater and subterranean navigation [15], [35]. Unlike vision and LiDAR sensors that are impacted by low lighting and smoke, the radar system is robust in both extreme weather conditions in the outdoor and visually degraded indoor environments, while uniquely providing relative velocity information [9].

There exists a plethora of work [3], [30], [1] that explores the spinning frequency modulated continuous wave (FMCW) radar pioneered by the Oxford RoboCar dataset [23] with success in automotive applications. In contrast, there are still relatively few works that consider the potential of the medium range imaging radar. Compared to spinning FMCW radars used in the automotive settings, medium range system-on-chip imaging radars have features that are appealing especially for indoor environments. They provide 3D information, are lightweight, less expensive, have a smaller form factor, and require lower power to operate. This makes them suitable for vehicles with limited carrying and power capacity that are commonly seen in indoor scenarios [16].

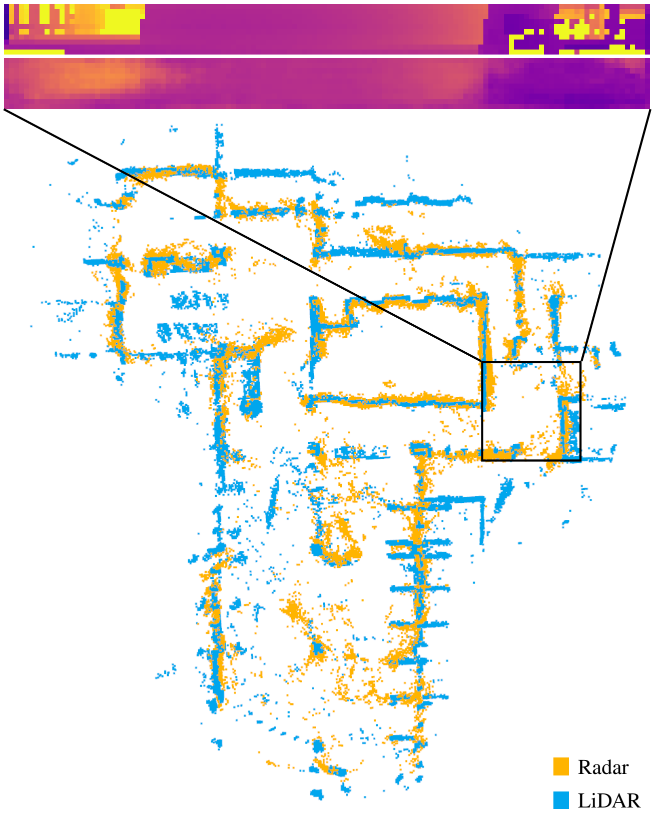

Fig. 1. An example of reference LiDAR depth map (top row) and inferred radar depth map (second row), with their view point marked on the map (black bounding box). Occupancy mapping using inferred radar depth map (yellow), compared with that of LiDAR depth map (blue) demonstrates overall ability to generalize in the indoor scenario. Ceiling and floor removed for visual clarity.

Imaging radar however, has distinct characteristics that make it difficult to work with. Beyond the noisy measurements typically observed in spinning radars, since the angular information of detected targets is resolved through antenna arrays, the antenna placement affects the resolution and accuracy of the angular dimensions asymmetrically. Fig. 2 is an example comparison between the radar heatmap obtained post analog to digital converter (ADC) and the LiDAR measurement. It illustrates that the radar is not only unable to resolve azimuth direction with clarity, but also provides no clear association along the elevation axis: all range-azimuth slices at different elevations are nearly indistinguishable.

Classically, the next phase is a target detection approach through a detector such as the constant false alarm rate (CFAR) [8] or its variants. These methods filter the heatmap into a sparse set of point targets by detecting peaks based

This work was partially supported by Amazon Lab126.

The authors are with the Robotics Institute, Carnegie Mellon University (CMU), Pittsburgh, PA 15213, USA. {ruoyangx, weidong, akashsharma, kaess}@andrew.cmu.edu

[1] https://github.com/rpl-cmu/learned-depth-imaging-radar

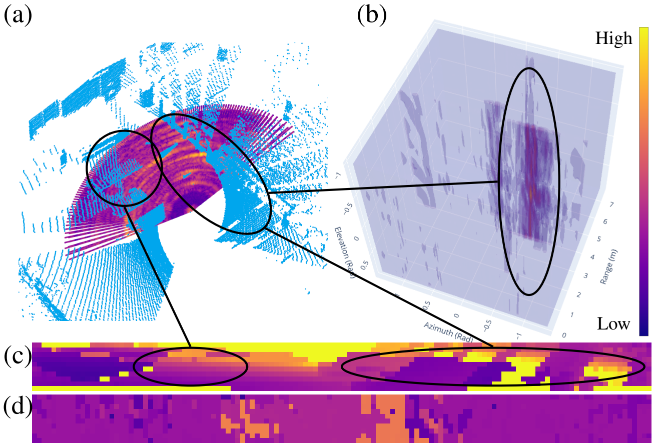

Fig. 2. Illustration of noise in elevation and azimuth axis. Connected circles denote correspondence between figures. (a) Radar intensity volume backprojected to Euclidean space (plasma colormap), and corresponding LiDAR scans in blue. Radar measurements only corresponding to the center elevation slice in the intensity volume are visualized. (b) Radar intensity volume in native spherical coordinate system. Consistent measurement in elevation axis shows the inability to resolve elevation angle with clarity. Elongated region along azimuth axis shows the difficulty in resolving accurate azimuth angle. (c) Reference LiDAR depth map. (d) Radar depth map obtained from the depth reading at highest radar returns along the range axis on the original intensity volume.

on an estimate of local noise. While this reduces the dimensionality of the observation drastically, a substantial amount of potentially valuable information is thrown out. Target detection also requires manual tuning of parameters to identify targets correctly. In contrast, we utilize information in the radar intensity volume to produce a dense estimate that addresses the aforementioned issues.

In particular, we present a learning-based method that regresses a dense depth map from the intensity volume of 4D imaging radar data. Our main contributions are two-fold:

1) A learned method for inferring a depth map for single frame imaging radar in a generalized indoor scene using LiDAR supervision.
2) We demonstrate the effectiveness of our method through the downstream tasks of 3D occupancy mapping, and body frame velocity estimation. An example mapping result is shown in Fig. 1.

Additionally, our method shows the potential for presenting 3D imaging radar in the form of more popular sensors in robotics. Following a literature review in Section II, we present and discuss our method in Section III. Finally, we provide experiments results of both depth map regression, and mapping in Section IV.

II. Related Work

A. Radar Imaging Systems

There exist several well-established close-range mmWave imaging radar systems and datasets [11], [24], [29] They demonstrate high resolution radar imaging, however, they only operate over very short ranges, or require a bulky sensor setup and radar absorbing materials in the background, making them impractical for navigation tasks. Another approach is to utilize motion to simulate a synthetic aperture radar [28], [26], [25]. While the results are promising, especially in the case of *MilliPoint* where accurate elevation information is also available as an output, one shortcoming of these methods is that the resulting imaging depends on multiple observations, which poses difficulty for tasks when accurate motion information is not readily available.

Recently, works have leveraged deep learning in the context of mmWave radar data. One line of work addresses radar-camera fusion using automotive spinning FMCW radars [20], [10], [21], while others show reasonable success in learning human poses from radio frequency (RF) signals [34], [33] or filtering information from classical detectors [19]. Most notable is the application of a conditional generative adversarial network (cGAN) for imaging radar depth estimation proposed in [12], [27]. These methods achieve impressive results for depth map inference from radar intensity volumes, however both methods are specialized for single class single object estimation, and the method scales poorly when multiple instances of the trained object are seen in the input image. Their ability to generalize to more generic scenes is questionable.

B. Learned Depth Estimation from Images

Learning-based methods are capable of fitting the function between camera images and their corresponding depth maps [2], [18]. In terms of formulation, our work is closely related to depth refinement since the radar intensity volume can be considered as a coarse estimate of depth. However, we find that depth refinement is typically a subtask in monocular or multi-view stereo depth estimation [32], [31], where the objective is to remove and smoothen artifacts around contours of similar depths. Such refinement processes have access to RGB images that innately encode high-frequency features such as boundaries and edges, providing excellent local information for regression and smoothing of pixel depth, unlike radar. Additionally, due to the noisy nature of resolving angle of arrival from targets in the imaging radar, it is common for multiple peaks to reside along the same single beam of direction due to the noise of multiple adjacent targets. It should be noted that these are not multi-path reflections. Such a characteristic sensor model complicates leveraging structure, such as the association of features or recognition of structures,

It is also well-known that radar has very different penetration and reflection characteristic from laser and camera sensors [9]. Therefore, ground truth generation for radar is not a straightforward process for learning, and is either difficult, reliant on synthetic data, or assumes that laser measurements are the ground truth.

C. Mapping with Radar

Limited range and angular resolution, multi-path reflections, and sparse measurements present challenges for 3D mapping using radar. One approach uses handcrafted prior knowledge of environment geometry, such as corridor width, to filter radar measurements [5]. With accurate odometry, and

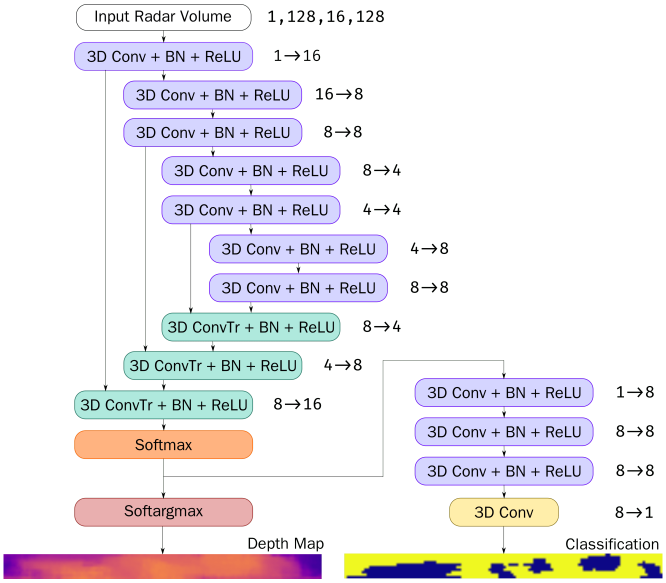

Fig. 3. 3D convolutional network architecture for depth regression and out-of-range classification. The network has a U-net structure for the primary section of the network. The classification has kernel and stride sizes designed to reduce the range dimension to 1.

good knowledge of the environment, occupancy grids can be built from sparse targets, however these assumptions are rarely true when operating in a visually degraded environment.

Another approach to tackle mapping is to learn a deep model that generates a dense map given a set of radar detections. *MilliMap* [22] stitches multiple frames of scans from a single-chip mmWave radar together in the form of 2D occupancy grids, and uses a conditional GAN to complete the map. While it shows promising results, collecting the large amount of training data required is very costly.

Recent work has shown significant progress and potential in using mmWave imaging radar for robot perception, yet the challenge of extracting dense 3D information from a single frame of radar data has not been sufficiently tackled. This motivates our work to create a system that allows for single frame 3D perception for imaging radar, enabling application of imaging radar for robot navigation tasks.

III. Method

We propose a 3D convolution based supervised regression model for the task of depth estimation. This section describes the detailed architecture of the proposed network and the data representation used for input and supervision.

A. Data Representation

The measurements from mmWave radar has four dimensions: range, azimuth, elevation, and with the last dimension consists of intensity and velocity. From this 4D heat-volume, we use the 3D per-voxel intensity as our input. Since radar measures environment by range and angle, this representation is natively in the 3D spherical coordinate system of (r, ϕ, θ) (range, azimuth, elevation). We assume that laser measurements are a good proxy for the true radar measurements, even though mmWave radar has different penetrative characteristics than laser sensors.

There are several ways to formulate radar-laser supervision. One formulation would be to backproject radar points into their Euclidean coordinate representation and then classify individual points. [4]. However due to the spherical coordinate system, adjacent points in the intensity volume grow farther apart with increasing range, which makes it difficult to capture local relationships which should be invariant to the range value. This is made worse with the spare resolution. Second, it is also possible to construct an occupancy volume in the same spherical coordinate system from LiDAR measurements. Adopting a depth map formulation retains local relationships, and is a simpler representation of the sensor model. Compared to the volumetric or unprojected points, the depth map representation loses the ability to distinguish between multiple detections along the same angular orientation, that could be caused by multi-path or behind-the-wall detections. However in the context of indoor mapping, with the detection range of medium range imaging radar at multiple-input multiple-output (MIMO) mode at less than $8m$, this loss of representation does not lose significant information.

The ground truth depth maps are generated by projecting LiDAR scans into a cylindrical depth map. To avoid choppy images caused by calibration issues, we utilize the formulation introduced in [6]. The radar intensity volume is further cropped to have a similar vertical field of view (FoV) as LiDAR.

B. System Overview

The system comprises primarily of two sections, a depth map regression model and an out-of-range pixel classification model. This pixel classification model accounts for two shortcomings in the radar based regression: 1) a significant number of pixels in the depth map are often out of the radar detection range, due to the relatively short range of imaging radar and 2) large noise in the angular coordinates requires the use of local information (convolution) to discriminate a false return due to targets in adjacent region from a true radar return. The network architecture is illustrated in Fig. 3.

C. Depth Map Estimation

The depth map regression takes heavy intuition from cell-averaging CFAR, which could be viewed as a rectangular filter. However, the parameter tuning for CFAR is time-consuming and hard to evaluate due to the sparse spatial resolution. We recognize the similarity between CFAR and convolution, and the success in learning-based reconstruction, especially in [32] where the feature points are collected into a cost volume trained to generate a probability volume for the estimated depths. We adopt a probabilistic view where the raw intensity volume observed are noisy measurements from which, an underlying true depth can be estimated.

With that intuition, we use a multi-scale 3D convolutional backend for depth regression. We use skip connections to link earlier convolutional blocks to the deconvolution layers, batch normalization and ReLU activations for the deconvolution layers. The last deconvolution layer outputs a 1-channel 3D volume.

We define range, azimuth, and elevation angle bin $\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\phi}$ of the radar measurement volume and note the original intensity volume as $\mathbf{V}_I$ and output volume at the last deconvolution layer $\mathbf{P}$, $\{\mathbf{V}_I, \mathbf{P}\} \in \mathbb{R}^{|\mathbf{r}|\times|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$. To preserve differentiability, we compute the 2D image $\mathbf{I}_r \in \mathbb{R}^{|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$ through soft argmax that computes the expected value given the distribution

$$\mathbf{I}_r(\theta, \phi) = \sum_{r=0}^{|\mathbf{r}|} \mathbf{r}(r)\mathbf{P}(r, \theta, \phi) \qquad (1)$$

D. Out-of-range Invalid Points

The true radar depth map is limited by the range of space covered by $\mathbf{P}$. Therefore there are depth values in $\mathbf{I}_r$ that should be returned as out of bounds. While we do observe a positive correlation between the standard deviation of probability along the depth axis in $\mathbf{P}$ for the out-of-range pixels, the high false positive and true negative rate requires a more intelligent processing method. Other methods such as photometric confidence proposed in [32], which calculates the confidence in prediction based on the surrounding values of the maxima, were also experimented on to no significant success.

The task here is to reduce $\mathbb{R}^{|\mathbf{r}|\times|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$ to $\mathbb{R}^{|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$ while maintaining invariance to the position of values along the depth dimension. We propose to use consecutive 3D convolutional blocks with strides and kernel sizes that incrementally reduce the dimension along the depth axis while maintaining the dimensions for azimuth and elevation. Each block has eight channels with batch normalization and ReLU as activation function. The last layer outputs an image for pixel-wise classification $\mathbf{I}_c \in \mathbb{R}^{2\times|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$.

E. Loss Function

The loss function considers both the classification of out-of-range pixels and depth regression. We only consider in-range pixels for the depth estimation task.

$$Loss = l_{\text{BCE}}(\mathbf{I}_m, \mathbf{I}_c) + \sum_{p \in \mathbf{I}_l < r_{\max}} \psi\left(\mathbf{I}_l(p) - \mathbf{I}_r(p)\right) \qquad (2)$$

Here $p \in \mathbf{I}_l < r_{max}$ denotes the set of pixels that are within maximum range of $\mathbf{r}$, $\mathbf{I}_l$ the LiDAR ground truth, $\mathbf{I}_m \in \mathbb{R}^{2\times|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$ the ground truth map for in-range and out-of-range pixels. ψ denotes Huber robust cost [14] function to account for situations where radar measurements detect significantly different objects from LiDAR. l_{BCE} is the binary cross entropy loss.

IV. Experiments and Evaluation

We evaluate our method on the publicly available mmWave imaging radar dataset ColoRadar [16]. The dataset provides data from an IMU, LiDAR, a Texas Instruments (TI) cascaded imaging radar (AWR2243) operating in MIMO mode, and a single chip radar (AWR1843). The dataset contains sequences through both indoor and outdoor environments as well as ground truth trajectory generated from LiDAR-inertial SLAM methods.

Specifically, we train and test in the *ec_hallways* and *arpg_lab* sequences. These two scenarios are representative of an ordinary building as they travel through corridors and large rooms. The sensor rig performs quadrotor-like motion during data recording. We train the network on sequence 2 of *ec_hallways* and test on the rest of the sequences. Sequence 3 of *ec_hallways* is omitted due to a ~20s duration of dropped radar frame in the middle of the run. Training is performed on an NVIDIA RTX2070S with a batch size of 16 for 150 epochs using Adam optimizer.

A. Depth Map Estimation and Out-of-Range Classification

1) Depth Estimation: We qualitatively show our outputs in Fig. 4, where we compare the LiDAR and inferred radar depth map, the unprojected points of both LiDAR and radar depth map, and the original $\mathbf{V}_I$. Our method generates depth maps that are visually consistent with LiDAR ground truth. As a baseline, our method can capture visually significant peaks in the original radar volume such as pillars marked in the green bounding boxes. Most significantly, it can capture ceilings and floors that are almost unperceivable from the original radar volume, as marked by the two elongated blue bounding boxes on the depth map.

We use metrics commonly seen in the monocular depth estimation literature to evaluate the depth map results [7]. We provide a brief summarization in Table I. The quantitative results are summarized in Table II. For the metrics, ↑ shows higher is better, and vice versa.

arpg_lab runs score lower overall performance than *ec_hallways*. This is expected since there could be overlaps between the training sequence 2 and the rest of *ec_hallways* sequences. However, since *arpg_lab* catches up to the performance of *ec_hallways* when threshold δ value increases from 1.25 to 1.25^2, the degradation in performance is relatively local.

2) Out-of-Range Classification: In this section, we compare the quantitative results for the out-of-range pixel classi-

TABLE I

Depth Evaluation Metrics

Abs rel: $\frac{1}{\lvert T\rvert}\sum_{y\in T}\frac{\lvert\tilde{y}-y^*\rvert}{y^*}$	RMSE: $\sqrt{\frac{1}{\lvert T\rvert}\sum_{y\in T}\lvert\lvert\tilde{y}-y^*\rvert\rvert^2}$
Sqr rel: $\frac{1}{\lvert T\rvert}\sum_{y\in T}\frac{\lvert\lvert\tilde{y}-y^*\rvert\rvert^2}{y^*}$	Thr: $\%\ y \ni \max(\frac{\tilde{y}}{y^*}, \frac{y^*}{\tilde{y}}) = \delta < thr$

TABLE II

Quantitative Results on Depth Estimation

	Error (↓)			δ Threshold, (↑)		
	Abs Rel	Sqr Rel	RMSE	1.25	1.25^2	1.25^3
EC 0	0.2057	0.3197	1.1216	0.7004	0.881	0.9491
EC 1	0.2097	0.2878	0.9916	0.6996	0.8929	0.9593
EC 4	0.2407	0.3895	1.2138	0.6439	0.8441	0.9299
Arpg 0	0.2646	0.4266	1.3312	0.5447	0.8084	0.92
Arpg 1	0.2664	0.4324	1.3222	0.5555	0.8043	0.9118
Arpg 2	0.2723	0.4362	1.3391	0.5332	0.7864	0.9098
Arpg 3	0.2635	0.3989	1.2288	0.5691	0.8196	0.9267
Arpg 4	0.2595	0.4077	1.2612	0.573	0.8246	0.924

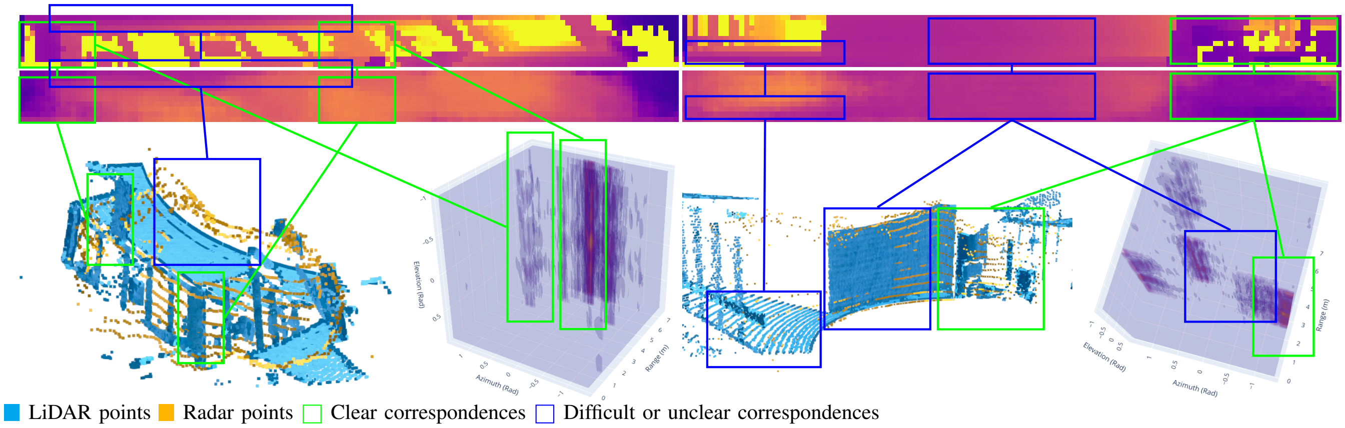

Fig. 4. Qualitative performance on indoor scenarios. The first row shows LiDAR depth maps, the second row shows inferred radar depth maps. Brighter is farther. The third row contains unprojected points and raw radar intensity volume. In the unprojected points figure, LiDAR points are colored blue, radar points are colored yellow. The original density of LiDAR points are used for better visualization of the scenario. In the radar intensity volume, brighter indicates higher intensity. Bounding boxes indicates correspondence between images: green bounding boxes show estimations that are visible in the raw radar volume; blue bounding boxes indicate environment features that are difficult to perceive in the raw radar volume. The figure shows that our method is able to generate visually accurate results, and capture floors and ceilings that are barely visible in the original intensity volume.

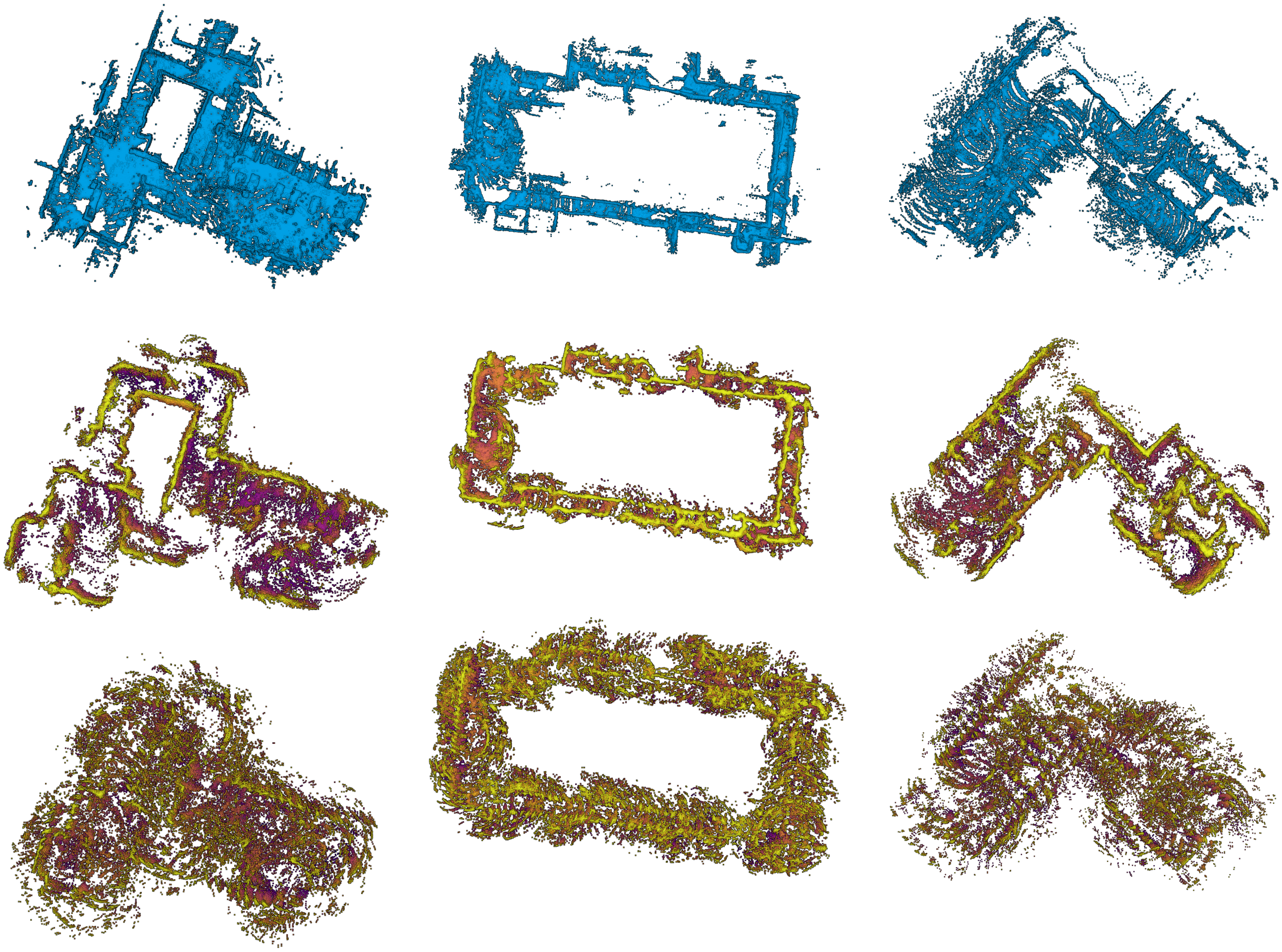

Fig. 5. From top to bottom: Mapping using LiDAR depth map limited to radar FoV; Mapping using inferred depth estimation of radar; Mapping using CFAR but only the first peak along each beam of direction. All of the maps have ceiling removed naïvely by thresholding over $0.8m$; the floors exist at around $-0.8m$. For the radar maps, darker denotes lower z value. Therefore brighter color usually denotes walls, and large portion of dark region indicates floors. Mapping using CFAR results in noisier maps due to the sparse nature of the detector and the false positive detections that appear independent of adjacent structures.

fication. This process is analogous to the depth map filtering process in many learning-based multi-view reconstruction methods. We compare our method against the photometric confidence method described in [32], as well as standard deviation for a conventional statistical measure. F-score is used to evaluate the performance. The quantitative results are summarized in Table III. Our 3D convolution based method achieved improved F-scores due to its ability to utilize local information to determine if a pixel is truly out of range.

3) Failure Cases: Fig. 6 shows typical failure cases of our method. In the red-bounding box, LiDAR measurements penetrated through glass and registered the wall behind the glass, while radar has its measurements absorbed by the glass, resulting in a noisy output that cannot be rejected through the classification model since the output does not match the typical out-of-range characteristics. While failing to register the glass presents a potential danger for navigation tasks from the LiDAR side, radar also fails to detect the glass with clarity. It would be an interesting future work to enable the current system to understand different material properties as [22] did. The blue bounding box is a situation where our methods generated significantly noisier estimates. These situations typically happen when viewing doors from close-range. We suspect it is caused by a combination of noisy close range measurements and the complex reflection path formed by the angles of door frames when in close range.

B. Mapping Using Radar

In this section, we provide qualitative results for the overall depth map estimation through 3D occupancy mapping using ground truth poses provided in the dataset. Occupancy mapping is performed using OctoMap [13] with cell resolution $0.1m$ and default parameters for occupancy updates. The ground truth is constructed using the LiDAR depth maps created earlier for depth estimation supervision. We also compared our method to the classical CFAR detector. To adapt to the task of mapping, only the first peak detected along each beam of direction is taken as the valid detection. CFAR is unavailable for quantitative comparison in the earlier Section IV-A.1 due to its sparsity: our method consistently produces measurements of more than 1k points per frame, while CFAR produces less than 80. We show three mapping results from *ec_hallway* sequences. All of the ceilings of the mapped scene have been removed for better visualization. The result are shown in Fig. 5. Our method successfully captures a vast majority of the structural features, with limited coverage of floors. Mapping using CFAR also results in traces of structural geometry, however the map is much noisier and less usable for robot navigation tasks. Additionally, CFAR has a hard time identifying the correct elevation for a target, which resulted in incorrect spherical surfaces directly under the xy-plane of the map.

TABLE III

QUANTITATIVE RESULTS ON PIXEL CLASSIFICATION

	F-Score (↑)		
	Seg	Pho.Conf	std
EC 0	0.7893	0.6294	0.4180
EC 1	0.8311	0.6655	0.3963
EC 4	0.7957	0.6238	0.4088
Arpg 0	0.8204	0.6466	0.4924
Arpg 1	0.8237	0.6513	0.4688
Arpg 2	0.8218	0.6346	0.4832
Arpg 3	0.8314	0.6542	0.4853
Arpg 4	0.8350	0.6716	0.4765

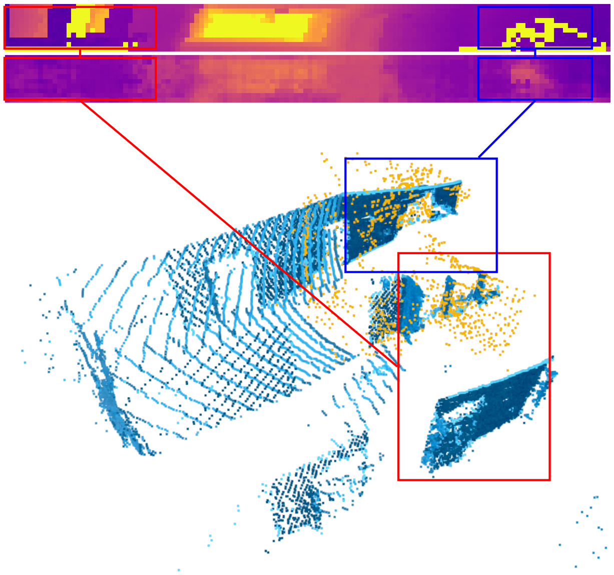

Fig. 6. Examples where the system performed poorly. From top to bottom: LiDAR depth map, radar depth map, and depth map unprojected into 3D space. In the unprojected points figure, LiDAR points are painted blue and unprojected radar points are painted yellow. Connected boxes show correspondence. The red and blue bounding box shows situations where our method fails. The yellow structure inside the blue bounding box in LiDAR depth map is from dangling wires from the sensor rig. Looking from other view points, the overshadowed part is a door.

C. Application to Body Frame Velocity Estimation

Additionally, we show that our learning based depth estimation captures, to some degree, the "real" geometry by estimating the body velocity through indexing depth map into the velocity volume provided in the 4D radar data.

For the radar measurement $\mathbf{V}_v \in \mathbb{R}^{|\mathbf{r}|\times|\boldsymbol{\theta}|\times|\boldsymbol{\phi}|}$ where $\mathbf{V}_v(r,\theta,\phi)$ measures the velocity of the target relative to the sensor along the beam that crosses the sensor origin, with orientation defined by (θ,ϕ). Note that the velocity volume contains noise large enough that it cannot be used to filter invalid voxels in $\mathbf{V}_I$ even when given a good body velocity estimate. However, when given an estimate of sensor frame velocity $\mathbf{v}_s$, the error can be calculated as the following per [17], where $\mathbf{p}_v$ are valid radar detections:

$$e = \sum_{\{r,\theta,\phi\}\in\mathbf{p}_v} \mathbf{V}_v(r,\theta,\phi) + \mathbf{v}_s^\top \left(\frac{\mathbf{t}(r,\theta,\phi)}{||\mathbf{t}(r,\theta,\phi)||}\right) \tag{3}$$

$$\mathbf{t}(r,\theta,\phi) = [x,y,z]^\top = r\begin{bmatrix}\cos(\phi)\cos(\theta)\\ \cos(\phi)\sin(\theta)\\ \sin(\phi)\end{bmatrix}$$

TABLE IV

RMSE for Body Frame Velocity Estimation

	RMSE (m/s) ($\downarrow$)		
	Seq #0	Seq #1	Seq #4
v_x	0.5674	0.5915	0.4923
v_y	0.1796	0.2525	0.1838
v_z	0.2623	0.3406	0.2851

Without loss for generality, we assume that there are enough measurements for non-degeneracy. We directly solve for sensor frame velocity through

$$\tilde{\mathbf{v}}_s = \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^3} \sum_{\{r,\theta,\phi\}\in\mathbf{p}_v} \mathbf{V}_v(r,\theta,\phi) + \mathbf{v}^\top \left(\frac{\mathbf{t}(r,\theta,\phi)}{||\mathbf{t}(r,\theta,\phi)||}\right), \tag{4}$$

in the form of $\arg\min_x ||Ax - b||^2$ and the covariance is calculated as:

$$\Sigma = \left(A'^\top A'/||\mathbf{p}_v||\right)^{-1}, A' = A/(v_{\text{resoln}}/\sqrt{12}). \tag{5}$$

We can now assemble $\mathbf{p}_v$ by bitwise masking $\mathbf{I}_r$ with $\mathbf{I}_c$ and calculate the sensor frame velocity. We compare radar-inferred body frame velocity with inferred ground truth velocity through RMSE in Table IV.

In all of the sequences, the primary motion is around $1.2m/s$ in the y-axis with fluctuations. Radar experiences the largest error along the x-axis due to the poorer angular resolution as it gets closer to the lateral axis. We argue that since velocity is not part of the learning problem formulation, yet the velocity obtained from the learning outcome closely matches the true body velocity, it implies there exists some degree of learning of true geometry in obtaining dense radar measurements.

V. Conclusion

In this work we present a learning-based method for imaging radar perception. We formulate the output of radar measurements as cylindrical depth maps with LiDAR supervision. A pixel-wise classification module is created to filter out out-of-range measurements. We test our method on a publicly available mmWave radar dataset and show visually meaningful results. While the raw ADC measurement from radar still encode much information such as material property, our work shows the potential to convert the noisy intensity volume of 4D radar measurements into a standard depth map formulation with acceptable fidelity without any fine-tuning. This conversion shows the possibility of a unified radar data representation for navigation. In future work, we plan to explore techniques that enables imaging radar to be a more capable standalone sensor for navigation, as well as the ability to understand material property.

References

[1] R. Aldera, D. De Martini, M. Gadd, and P. Newman, "What Could Go Wrong? Introspective Radar Odometry in Challenging Environments," in *IEEE Intelligent Transportation Systems (ITSC) Conference*, Auckland, New Zealand, October 2019.

[2] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *AAAI*, 2019.

[3] S. H. Cen and P. Newman, "Radar-only ego-motion estimation in difficult settings via graph matching," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 298–304.

[4] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2D car detection in radar data with pointnets," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE Press, 2019, p. 61–66. [Online]. Available: https://doi.org/10.1109/ITSC.2019.8917000

[5] S. Dogru and L. Marques, "Using radar for grid based indoor mapping," in *2019 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2019, pp. 1–6.

[6] W. Dong, K. Ryu, M. Kaess, and J. Park, "Real-time registration and reconstruction with cylindrical lidar images," 2021.

[7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2366–2374.

[8] H. M. Finn and R. S. Johnson, "Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimates," *RCA Rev*, vol. 29, pp. 414–464, 1968.

[9] K. Garcia, M. Yan, and A. Purkovic, "Robust traffic and intersection monitoring using millimeter wave sensors," Texas Instruments, Tech. Rep., 2018.

[10] S. Gasperini, P. Koch, V. Dallabetta, N. Navab, B. Busam, and F. Tombari, "R4Dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes," in *2021 International Conference on 3D Vision (3DV)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2021, pp. 751–760. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/3DV53792.2021.00084

[11] M. T. Ghasr, M. J. Horst, M. R. Dvorsky, and R. Zoughi, "Wideband microwave camera for real-time 3-D imaging," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 1, pp. 258–268, 2017.

[12] J. Guan, S. Madani, S. Jog, S. Gupta, and H. Hassanieh, "Through fog high-resolution imaging using millimeter wave radar," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 461–11 470.

[13] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at https://octomap.github.io. [Online]. Available: https://octomap.github.io

[14] P. J. Huber, "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73 – 101, 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177703732

[15] H. Johannsson, M. Kaess, B. Englot, F. Hover, and J. Leonard, "Imaging sonar-aided navigation for autonomous underwater harbor surveillance," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4396–4403.

[16] A. Kramer, K. Harlow, C. Williams, and C. Heckman, "ColoRadar: The direct 3D millimeter wave radar dataset," vol. abs/2103.04510, 2021. [Online]. Available: https://arxiv.org/abs/2103.04510

[17] A. Kramer, C. Stahoviak, A. Santamaria-Navarro, A.-a. Aghamohammadi, and C. Heckman, "Radar-inertial ego-velocity estimation for visually degraded environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 5739–5746.

[18] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 239–248.

[19] C.-H. Lin, Y.-C. Lin, Y. Bai, W.-H. Chung, T.-S. Lee, and H. Huttunen, "DL-CFAR: A novel CFAR target detection method based on deep learning," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–6.

[20] J.-T. Lin, D. Dai, and L. V. Gool, "Depth estimation from monocular images and sparse radar data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 233–10 240.

[21] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, "Radar-camera pixel depth association for depth completion," in *IEEE Conference on Computer Vision and Pattern Recog-*

nition, CVPR 2021, virtual, June 19-25, 2021. Computer Vision Foundation / IEEE, 2021, pp. 12 507–12 516.

[22] C. X. Lu, S. Rosa, P. Zhao, B. Wang, C. Chen, J. A. Stankovic, N. Trigoni, and A. Markham, "See through smoke: Robust indoor mapping with low-cost mmWave radar," in *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2020.

[23] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

[24] B. Mamandipoor, G. Malysa, A. Arbabian, U. Madhow, and K. Noujeim, "60 GHz synthetic aperture radar for short-range imaging: Theory and experiments," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 553–558.

[25] K. Qian, Z. He, and X. Zhang, "3D point cloud generation with millimeter-wave radar," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, dec 2020. [Online]. Available: https://doi.org/10.1145/3432221

[26] C. Stetco, B. Ubezio, S. Mühlbacher-Karrer, and H. Zangl, "Radar sensors in collaborative robotics: Fast simulation and experimental validation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 452–10 458.

[27] 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)*, 2021, pp. 1–8.

[28] 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 *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4304–4310.

[29] S. Wei, Z. Zhou, M. Wang, J. Wei, S. Liu, J. Shi, X. Zhang, and F. Fan, "3DRIED: A high-resolution 3-D millimeter-wave radar dataset dedicated to imaging and evaluation," *Remote Sensing*, vol. 13, no. 17, p. 3366, Aug 2021. [Online]. Available: http://dx.doi.org/10.3390/rs13173366

[30] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably unknown: Deep inverse sensor modelling radar," 2019.

[31] S. Yan, C. Wu, L. Wang, F. Xu, L. An, K. Guo, and Y. Liu, "DDRNet: Depth map denoising and refinement for consumer depth cameras using cascaded CNNs," in *ECCV*, 2018.

[32] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," *European Conference on Computer Vision (ECCV)*, 2018.

[33] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7356–7365.

[34] M. Zhao, Y. Liu, A. Raghu, H. Zhao, T. Li, A. Torralba, and D. Katabi, "Through-wall human mesh recovery using radio signals," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 112–10 121.

[35] S. Zhao, P. Wang, H. Zhang, Z. Fang, and S. Scherer, "TP-TIO: A robust thermal-inertial odometry with deep thermalpoint," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. [Online]. Available: https://arxiv.org/abs/2012.03455