2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Probably Unknown: Deep Inverse Sensor Modelling Radar

Rob Weston, Sarah Cen, Paul Newman and Ingmar Posner

***Abstract*— Radar presents a promising alternative to lidar and vision in autonomous vehicle applications, able to detect objects at long range under a variety of weather conditions. However, distinguishing between occupied and free space from raw radar power returns is challenging due to complex interactions between sensor noise and occlusion.**

To counter this we propose to learn an Inverse Sensor Model (ISM) converting a raw radar scan to a grid map of occupancy probabilities using a deep neural network. Our network is self-supervised using partial occupancy labels generated by lidar, allowing a robot to learn about world occupancy from past experience without human supervision. We evaluate our approach on five hours of data recorded in a dynamic urban environment. By accounting for the scene context of each grid cell our model is able to successfully segment the world into occupied and free space, outperforming standard CFAR filtering approaches. Additionally by incorporating heteroscedastic uncertainty into our model formulation, we are able to quantify the variance in the uncertainty throughout the sensor observation. Through this mechanism we are able to successfully identify regions of space that are likely to be occluded.

I. INTRODUCTION

Occupancy grid mapping has been extensively studied [1], [2] and successfully utilised for a range of tasks including localisation [3], [4] and path-planning [5]. One common approach to occupancy grid mapping uses an inverse sensor model (ISM) to predict the probability that each grid cell in the map is either *occupied* or *free* from sensor observations. Whilst lidar systems provide precise, fine-grained measurements, making them an obvious choice for grid mapping, they fail if the environment contains fog, rain, or dust [6]. Under these and other challenging conditions, FMCW radar is a promising alternative that is robust to changes in lighting and weather and detects long-range objects, making it well suited for use in autonomous transport applications.

However, two major challenges must be overcome in order to utilise radar to this end. Firstly, radar scans are notoriously difficult to interpret due to the presence of several pertinent noise artefacts. Secondly, by compressing information over a range of heights onto a dense 2D grid of power returns identifying occlusion becomes difficult. The complex interaction between occlusion and noise artefacts introduces uncertainty in the state of occupancy of each grid cell which is *heteroscedastic*, varying from one world location to another based on scene context, and *aleatoric* [7], inherent in radar data by way of the scan formation process.

In order to successfully reason about world occupancy, we posit that a model that is able to reason about scene context is essential. To this end, we formulate the problem of determining an ISM as a segmentation task, leveraging a deep network to learn the probability distribution of occupancy from raw data alone. This allows us to successfully determine regions of space that are likely to be occupied and free in light of challenging noise artefacts. Simultaneously, by explicitly modelling heteroscedastic uncertainty, we are able quantifying to quantify the latent uncertainty associated with each world cell arising through occlusion. Utilising approximate variational inference we are able to train our network using self-supervision relying on partial labels automatically generated from occupancy observations in lidar.

Authors are from the Oxford Robotics Institute (ORI)
{robw,sarah,pnewman,ingmar}@robots.ox.ac.uk

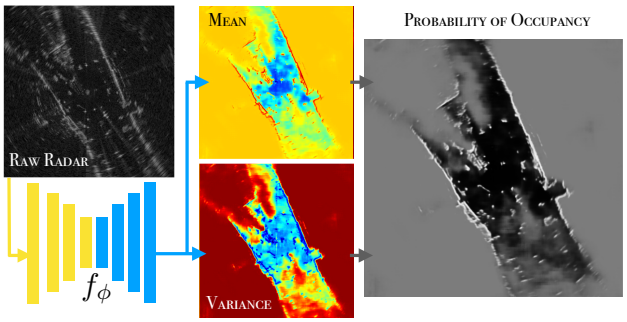


Fig. 1. Our network learns the distribution of occupancy from experience alone. By reasoning about scene context it is able to successfully identify regions of space that are likely to be occupied and free. The uncertainty associated with each grid cell is allowed to vary throughout the scene by predicting the noise standard deviation alongside the predicted logit of each grid cell. These are combined to generate a grid map of occupancy probabilities. The uncertainty predicted by our network can be used to successfully identify regions of space that are likely to be occluded.

We train our model on real-world data generated from five hours of urban driving and successfully distinguish between occupied and free space, outperforming constant false-alarm rate (CFAR) filtering in average intersection over union performance. Additionally we show that by modelling heteroscedastic uncertainty we are able to successfully quantify the uncertainty arising through the occlusion of each grid cell.

II. RELATED WORK

Inverse sensor models (ISMs) [1] are used to convert noisy sensor observation to a grid map of occupancy probabilities. For moving platforms, a world occupancy map can then be sequentially generated from an ISM, multiple observations, and known robot poses using a binary Bayes filter [**?**]. Using lidar data, ISMs are typically constructed using a combination of sensor-specific characteristics, experimental data, and empirically-determined parameters [8], [9], [10]. These human-constructed ISMs struggle to model challenging radar defects and often utilise limited local information to predict each cell's occupancy without accounting for scene context.

Instead, raw radar scans are often naively converted to binary occupancy grids using classical filtering techniques that distinguish between objects (or targets) and free space (or background). Common methods include CFAR [11] and static thresholding. However, both return binary labels rather

than probabilities, and neither is capable of addressing all types of radar defects or capturing occlusion. Additionally, the most popular approach, CFAR, imposes strict assumptions on the noise distribution and requires manual parameter tuning. In contrast, using deep learning methods, as first proposed by [12], allows the distribution of world occupancy to be learned from raw data alone, accounting for the complex interaction between sensor noise and occlusion through the higher level spatial context of each grid cell.

In order to capture uncertainty that varies from one grid cell to the next we incorporate heteroscedastic uncertainty into our formulation inspired by [7]. Our variational reformulation of [7] is closely related to the seminal works on variational inference in deep latent variable models [13], [14] and their extension to conditional distributions [15].

Drawing on the successes of deep segmentation in biomedical applications, [16] and vision [17] we reformulate the problem of learning an inverse sensor model as neural network segmentation. Specifically, we utilise a U-net architecture with skip connections [18]. In order to map from an inherently polar sensor observation to a Cartesian map we utilise Polar Transformer Units (PTUs) [19].

III. DEEP INVERSE SENSOR MODELLING IN RADAR

A. Setting

Let $\boldsymbol{x} \in \mathbb{R}^{\Theta \times R}$ denote a full radar scan containing Θ azimuths of power returns at R different ranges for each full rotation of the sensor. Partitioning the world into a $H \times W$ grid, $\boldsymbol{y} \in \{0,1\}^{H \times W}$ gives the occupancy state of each grid cell, where $\boldsymbol{y}^{u,v} = 1$ if cell (u,v) is *occupied* and $\boldsymbol{y}^{u,v} = 0$ if (u,v) is *free*. *Partial* measurements of occupancy $\hat{\boldsymbol{y}}$ are determined by combining the output of multiple 3D lidars and projecting the returns over a range of heights onto a 2D grid. In order to separate the region of space where no labels exist most likely as a consequence of full occlusion, from space that is likely to only be partially occluded or for which no labels exist due to a limited field of view of the lidar sensors, the observability state of each cell $\boldsymbol{o}^{u,v}$ is recorded as 0, 1 or 2 corresponding to *unobserved*, *observed* and *partially observed* space respectively. The full labelling procedure is described in Figure 2. This process is repeated for N radar-laser pairs to generate a data set $\mathcal{D} = \{\boldsymbol{x}^n, (\hat{\boldsymbol{y}}, \boldsymbol{o})^n\}_{n=1}^N$ of training examples from which we aim to learn an inverse sensor model $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}} \in [0,1]^{H \times W}$ such that $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}}^{u,v} = p(\boldsymbol{y}^{u,v} = 1 | \boldsymbol{x})$ gives the probability that cell (u,v) is occupied dependent on the *full* radar scan $\boldsymbol{x}$

B. Heteroscedastic Aleatoric Uncertainty and FMCW Radar

FMCW Radar is an inherently noisy modality suffering from speckle noise, phase noise, amplifier saturation and ghost objects. These conspire to make the distinction between occupied and free space notoriously difficult. A radar's long range as well as its ability to penetrate past first returns make it attractive but also challenging. In particular, a radar's capacity for multiple returns along an azimuth implies varying degrees of uncertainty depending on scene context:

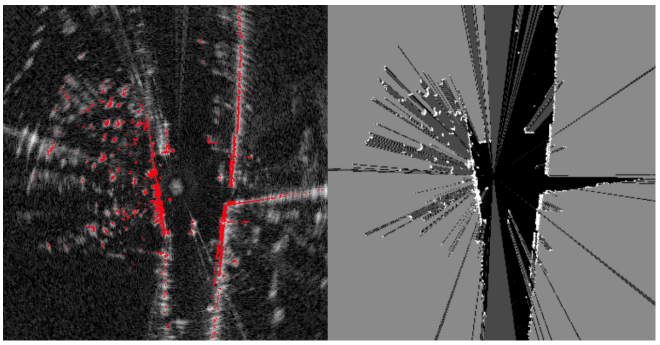


Fig. 2. Generated training labels from lidar. The image on the left shows the lidar points (red) projected into a radar scan $\boldsymbol{x}$ converted to Cartesain co-ordinates for visualisation. The right image shows the generated training labels. Any grid cell (u,v) with a lidar return is labelled as occupied $\hat{\boldsymbol{y}}^{u,v} = 1$ (white). Ray tracing along each azimuth, the space immediately in front of the first return is labelled as $\hat{\boldsymbol{y}}^{u,v} = 0$ (black), the space between the first and last return or along azimuths in which there is no return is labelled as *partially observed*, $\boldsymbol{o}^{u,v} = 2$, (dark grey) and the space behind the last return is labelled as *unobserved*, $\boldsymbol{o}^{u,v} = 0$, (light grey). Any space that is labelled as occupied or free is labelled as *observed*, $\boldsymbol{o}^{u,v} = 1$

the distinction between occupied and free space becomes increasingly uncertain as regions of space become partially occluded by objects. Examples of each of these problems are further explained in Figure 3. As such, high power returns do not always denote occupied and likewise, low power returns do not always denote free.

Uncertainties in our problem formulation depend on the world scene through a complex interaction between scene context and sensor noise, and are inherent in our data as a consequence of the image formation process. As such they are, heteroscedastic as they depend on scene context and aleatoric as they are ever present in our data [7]. In order to successfully determine world occupancy from an inherently uncertain radar scan we seek a model that explicitly captures heteroscedastic aleatoric uncertainty. By framing this problem as a deep segmentation task we leverage the power of neural networks to learn an ISM which accounts for scene context in order to determine – from raw data alone – *occupied* from *free* space in the presence of challenging noise artefacts. Simultaneously, as a result of our heteroscedastic uncertainty formulation we are also able to learn which regions of space are inherently uncertain because of occlusion.

C. Modelling Heteroscedastic Aleatoric Uncertainty

Instead of assuming that the uncertainty associated with each grid cell is fixed, as is typically assumed in standard deep segmentation approaches, by using a heteroscedastic model the uncertainty in each grid cell $\boldsymbol{\gamma}_\phi(\boldsymbol{x})$ is allowed to vary. This is achieved by introducing a normally distributed latent variable $\boldsymbol{z}^{u,v}$ associated with each grid cell [7] and predicting the noise standard deviation $\boldsymbol{\gamma}_\phi(\boldsymbol{x})$ alongside the predicted logit $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ of each each $\boldsymbol{z}^{u,v}$ with a neural network f_ϕ :

$$p_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_\phi(\boldsymbol{x}), \boldsymbol{\gamma}_\phi(\boldsymbol{x})\boldsymbol{I}) \quad (1)$$

$$[\boldsymbol{\mu}_\phi(\boldsymbol{x}), \boldsymbol{\gamma}_\phi(\boldsymbol{x})] := f_\phi(\boldsymbol{x}) \quad (2)$$

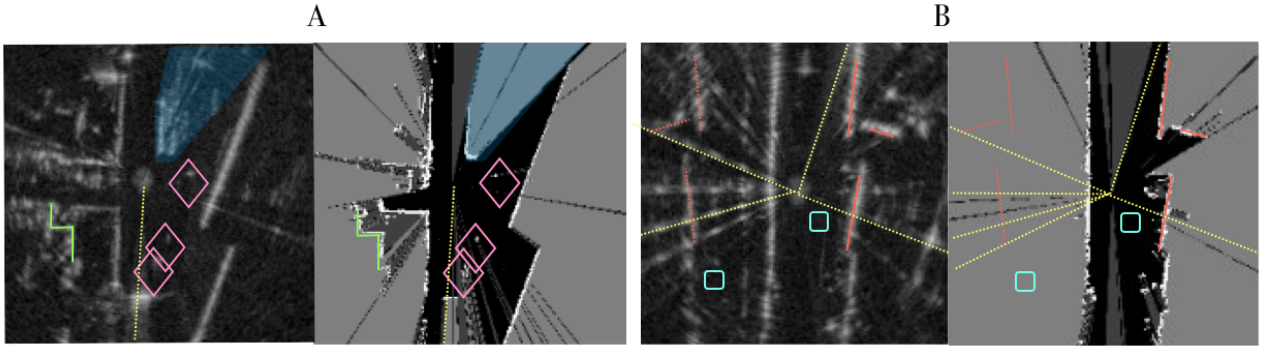


Fig. 3. Raw radar and the lidar ground truth. An ISM must be able to pick out faint objects, such as cars (pink diamonds), from the background speckle noise, in light of challenging noise artefacts such as saturation (yellow lines). In addition, an ISM must be able to determine which regions of space are likely to be occluded such as the space behind buses (highlighted blue) in light of almost identical local appearances (blue cyan boxes). Finally an ISM should be able to distinguish ghost objects (dotted orange) from true second returns (green lines).

Assuming a likelihood $p(\boldsymbol{y}^{u,v} = 1|\boldsymbol{z}^{u,v}) = \mathsf{Sigmoid}(\boldsymbol{z})$, the probability that cell $\boldsymbol{y}^{u,v}$ is occupied is then given by marginalising out the uncertainty associated with $\boldsymbol{z}$:

$$p(\boldsymbol{y}^{u,v}|\boldsymbol{x}) = \int p(\boldsymbol{y}^{u,v}|\boldsymbol{z}^{u,v}) p_\phi(\boldsymbol{z}^{u,v}|\boldsymbol{x}) d\boldsymbol{z}^{u,v} \tag{3}$$

Unfortunately the integral in (3) is intractable and is typically approximated using Monte-Carlo sampling and the reparameterization trick [7]. Instead, by introducing an analytic approximation in Section III-E we show that we can accurately and efficiently approximate (3) without resorting to sampling.

One final problem remains. We expect our model to be inherently uncertain in occluded space for which no lidar training labels are available. How do we train f_ϕ whilst explicitly encoding an assumption that in the absence of training labels we expect our model to be uncertain? In Section III-D we propose to solve this problem by introducing a normally distributed prior $p(\boldsymbol{z})$ on the region of space for which no training labels exist utilising the variational inference framework.

D. Training with Partial Observations

In order to encode an assumption that in the absence of training data we expect our model to be explicitly uncertain we introduce a prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\gamma}\boldsymbol{I})$ on the uncertainty associated with the occluded scene which our network reverts back to in the absence of a supervised training signal. To do this, we begin by treating $p_\phi(\boldsymbol{z}|\boldsymbol{x})$ as an approximate posterior to $p(\boldsymbol{z}|\boldsymbol{y})$ induced by the joint $p(\boldsymbol{z}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})$ where,

$$p(\boldsymbol{y}|\boldsymbol{z}) := \prod_{u,v} \mathsf{Bern}(\boldsymbol{y}^{u,v}|\boldsymbol{p}^{u,v}_{\boldsymbol{y}|\boldsymbol{z}}) \tag{4}$$

$$\boldsymbol{p}^{u,v}_{\boldsymbol{y}|\boldsymbol{z}} = p(\boldsymbol{y}^{u,v} = 1) = \mathsf{Sigmoid}(\boldsymbol{z}^{u,v}) \tag{5}$$

$$p(\boldsymbol{z}) := \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \gamma I) \tag{6}$$

$\mathsf{Sigmoid}$ and $\mathsf{Bern}(y|p) = p^y(1-p)^{1-y}$ denote the element-wise sigmoid function and Bernoulli distribution.

Next given a set of observations $\mathcal{D}$, we consider determining our parameters ϕ by maximising the variational lower bound,

$$\mathcal{L}(\phi; \mathcal{D}) = \sum_n \mathcal{L}^n(\phi) \tag{7}$$

$$\mathcal{L}^n(\phi) = \mathbb{E}_{p_\phi(\boldsymbol{z}|\boldsymbol{x}^n)}[\log p(\boldsymbol{y}^n|\boldsymbol{z})] - d_{kl}[p_\phi(\boldsymbol{z}|\boldsymbol{x}^n)||p(\boldsymbol{z})] \tag{8}$$

where d_{kl} denotes KL divergence. The first term in $\mathcal{L}^n(\phi)$ is the expected log-likelihood under the approximate posterior $p_\phi(\boldsymbol{z}|\boldsymbol{x})$ which, when optimised, forces the network to maximise the probability of each occupancy label $\boldsymbol{y}$. The second term forces $p_\phi(\boldsymbol{z}|\boldsymbol{x})$ towards the prior $p(\boldsymbol{z})$.

Crucially, by only evaluating the log-likelihood term in the labelled region of space and only evaluating the KL divergence term in occluded space, we are able to train our network to maximise the probability of our labels whilst explicitly encoding an assumption that in the absence of training labels we expect our network to be inherently uncertain. The latter is achieved by setting the prior to $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \gamma I)$ corresponding to an assumption that occluded space is equally likely to be free or occupied with a fixed uncertainty γ. We tested multiple values of γ and found that setting $\gamma = 1$ gave good results.

For a Gaussian prior and approximate posterior the KL divergence term can be determined analytically, whilst the expected log-likelihood is estimated using the reparameterization trick [14] by sampling $\boldsymbol{z}^l = \boldsymbol{\mu}_\phi(\boldsymbol{x}) + \boldsymbol{\gamma}_\phi(\boldsymbol{x}) \circ \boldsymbol{\epsilon}^l$ where $\boldsymbol{\epsilon}^l \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The expected log-likelihood is then approximated as $\mathbb{E}_{p_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{y}|\boldsymbol{z})] \approx -\frac{1}{L}\sum_l \left(\sum_{u,v} \mathbb{H}[\boldsymbol{y}^{u,v}, \boldsymbol{p}^{l,u,v}_{\boldsymbol{y}|\boldsymbol{z}}]\right)$ where $\mathbb{H}$ denotes binary cross entropy.

Finally our loss function becomes

$$\hat{\mathcal{L}}^n(\phi) = \frac{\bar{\omega}}{L} \sum_{l,u,v} \mathbb{I}(\boldsymbol{o}^{u,v} = 1) \mathbb{H}_\alpha[\hat{\boldsymbol{y}}^{n,u,v}, \boldsymbol{p}^{n,l,u,v}_{\boldsymbol{y}|\boldsymbol{z}}] + \sum_{u,v} \mathbb{I}(\boldsymbol{o}^{u,v} = 0) d_{kl}[p_\phi(\boldsymbol{z}^{u,v}|\boldsymbol{x}^n)||p(\boldsymbol{z}^{u,v})] \tag{9}$$

$$\hat{\mathcal{L}}(\phi; \mathcal{D}) = \frac{1}{N} \sum_n \hat{\mathcal{L}}^n(\phi) \tag{10}$$

where $\mathbb{I}$ denotes the indicator function which is equal to 1 if its condition is met and 0 otherwise.

In order to ensure that labelled and unlabelled data contribute equally to our loss we re-weight the likelihood term

with $\bar{\omega} = \omega HW/(\sum_{uv} \mathbb{I}(\boldsymbol{o}^{u,v} = 1))$. The hyper-parameter ω is used to weight the relative importance between our prior and approximate evidence. As there is also a significant class imbalance between occupied and free space we use weighted binary cross entropy $\mathbb{H}_\alpha$ where the contribution from the occupied class is artificially inflated by weighting each occupied example by a hyper-parameter α. Note that in the partially observed region $\boldsymbol{o}^{u,v} = 2$ there is no loss.

E. Inference

Given a trained model $p_{\phi_*}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_{\phi_*}(\boldsymbol{x}), \boldsymbol{\gamma}_{\phi_*}(\boldsymbol{x}))$ we now wish to determine the probability that each cell is occupied given input $\boldsymbol{x}$ by marginalising out the uncertainty associated with the latent variable $\boldsymbol{z}$:

$$p(\boldsymbol{y}^{u,v}|\boldsymbol{x}) := \int p(\boldsymbol{y}^{u,v}|\boldsymbol{z}^{u,v}) p_{\phi_*}(\boldsymbol{z}^{u,v}|\boldsymbol{x}) d\boldsymbol{z}^{u,v} \tag{11}$$

However, for likelihood $p(\boldsymbol{y}^{u,v}|\boldsymbol{z}^{u,v}) = \text{Sigmoid}(\boldsymbol{z}^{u,v})$ no exact closed form solution exists to this integral. Instead of resorting to Monte Carlo sampling we approximate the sigmoid function with a probit function and use the result that a Gaussian distribution convolved with a probit function is another probit function [20]. Following this analysis, it can be shown that,

$$p(\boldsymbol{y}^{u,v} = 1|\boldsymbol{x}) \approx \text{Sigmoid}\left(\frac{\boldsymbol{\mu}^{u,v}_{\phi_*}}{\boldsymbol{s}^{u,v}_{\phi_*}}\right) \tag{12}$$

where $\boldsymbol{s}^{u,v}_{\phi_*} = (1 + (\boldsymbol{\gamma}^{u,v}_{\phi_*}\sqrt{\pi/8})^2)^{1/2}$, $\boldsymbol{\mu}^{u,v}_{\phi_*} = \boldsymbol{\mu}^{u,v}_{\phi_*}(\boldsymbol{x})$ and $\boldsymbol{\gamma}^{u,v}_{\phi_*} = \boldsymbol{\gamma}^{u,v}_{\phi_*}(\boldsymbol{x}_*)$. This allows us to efficiently calculate $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}}$ as,

$$[\boldsymbol{\mu}_{\phi_*}, \boldsymbol{\gamma}_{\phi_*}] = f_{\phi_*}(\boldsymbol{x}) \tag{13}$$

$$\boldsymbol{s}_{\phi_*} = (1 + (\boldsymbol{\gamma}_{\phi_*}\sqrt{\pi/8})^2)^{1/2} \tag{14}$$

$$\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}} := \text{Sigmoid}\left(\frac{\boldsymbol{\mu}_{\phi_*}}{\boldsymbol{s}_{\phi_*}}\right) \tag{15}$$

Figure 4 shows $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}}$ approximated using (15) and Monte Carlo sampling for varying $\boldsymbol{\mu}_{\phi_*}$ and $\boldsymbol{\gamma}_{\phi_*}$. The Monte Carlo estimate takes of the order 10^4 samples to converge, whilst the analytic approximation provides a close approximation to the converged Monte Carlo estimate.

In equation (15) the predicted logit $\boldsymbol{\mu}_{\phi_*}$ can be thought of as giving the score associated with labelling an example as occupied; intuitively the higher the score the higher the probability that each cell is occupied. In contrast, the predicted deviation $\boldsymbol{\gamma}_{\phi_*}$ increases the entropy in the predicted occupancy distribution independent of the cells predicted score and captures uncertainties that cannot be easily explained by the predicted score alone.

IV. RESULTS

In this Section we show that our model, despite challenging noise artefacts, is able to successfully segment the world into occupied and free space achieving higher mean Intersection over Union (IoU) scores than cell averaging CFAR filtering approaches. In addition to this we are also able to explicitly identify regions of space that are likely to be occluded through the uncertainties predicted by our network. We provide several qualitative examples of our model operating in challenging real world environments and

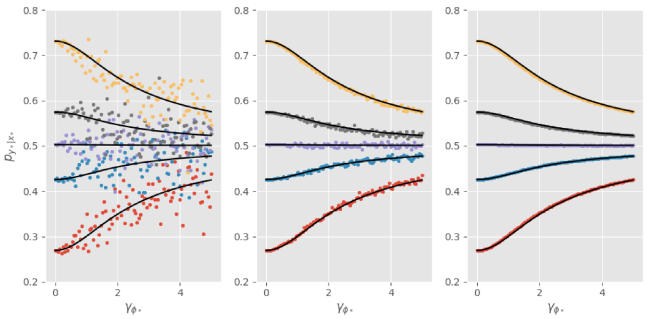


Fig. 4. Predicted occupancy probabilities $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}}$ as a function of predicted standard deviation $\boldsymbol{\gamma}_{\phi_*}$ using the analytic approximation given by (15) (black) vs Monte Carlo approximation with $L = 10^2$ (left), $L = 10^4$ (middle) and $L = 10^6$ (right) samples. Each colour corresponds to a different mean $\boldsymbol{\mu}_{\phi_*}$ with [yellow, grey, purple, blue, red] corresponding to means $[-1, -0.3, 0.01, 0.3, 1]$ respectively. It is seen that the MC estimate has high variance taking of the order 10^6 samples to converge to the analytic approximation. On the other hand the analytic approximation closely resembles the converged Monte Carlo estimate.

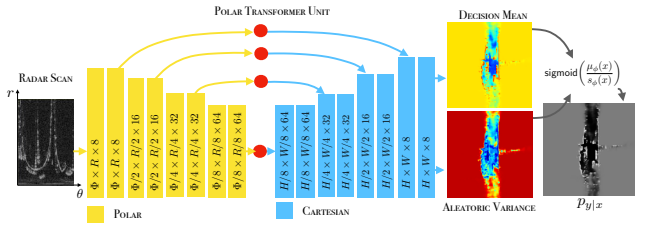


Fig. 5. Our network architecture takes in a polar radar scan $\boldsymbol{x} \in \mathbb{R}^{\Theta \times R}$ and maps it to Cartesian grids of mean utility $\boldsymbol{\mu}_\phi$ and aleatoric noise scale $\boldsymbol{s}_\phi = (1 + (\boldsymbol{\gamma}_\phi\sqrt{\pi/8})^2)^{1/2}$. Our network is composed of a polar (yellow) encoder and a Cartesian (blue) decoder. At each polar to Cartesian interface there is a polar transformer unit (red circle). Each blue rectangle corresponds to 2 convolutions followed by a max pool.

study the effects of our prior on our network output through an ablation study.

A. Experimental Set-Up

A Navtech CTS350x FMCW radar (without Doppler) and two Velodyne HDL32 lidars were mounted to a survey vehicle and used to generate over 78000 (90%) training examples and 8000 (10%) test examples from urban scenes. The output from the two lidars was combined from 0.7m below the roof of the vehicle to 1m above and projected onto a 600×600 grid, with a spatial resolution of 0.3m, generating a $180\text{m} \times 180\text{m}$ world occupancy map, following the procedure described in Section III-A. To account for differences in the frequency of our radar (4Hz) and lidar (10Hz) the occupancy map was ego-motion compensated such that the Cartesian map corresponds to the time stamps of each radar azimuth.

Figure 5 shows our network architecture in which a polar encoder takes the raw radar output and generates a polar feature tensor through repeated applications of 4×4 convolutions and max pooling before a Cartesian decoder maps this feature tensor to a grid of mean logits $\boldsymbol{\mu}_\phi(\boldsymbol{x}) \in \mathbb{R}^{H \times W}$ and standard deviations $\boldsymbol{\gamma}_\phi(\boldsymbol{x}) \in (0, \infty)^{H \times W}$ which are converted to a grid of probabilities through (15). Information is allowed to flow from the encoder to the decoder through skip connections, where polar features $\boldsymbol{u}$ are converted to Cartesian features $\boldsymbol{v}$ through bi-linear interpolation, with a fixed polar to Cartesian grid [19]. In all experiments we

TABLE I
COMPARING OUR APPROACH TO CLASSICAL DETECTION METHODS USING INTERSECTION OVER UNION

Method	Intersection over Union		
	Occupied	Free	Mean
CFAR (1D polar)	0.24	**0.92**	0.5
CFAR (2D Cartesian)	0.20	0.90	0.55
Static thresholding	0.19	0.77	0.48
Deep ISM (our approach)	**0.35**	0.91	**0.63**

trained our model using the ADAM optimiser [21], with a learning rate of 0.001, batch size 16 for 100 epochs and randomly rotated each input output pair about the origin, minimising the loss proposed in (9) with $L = 25$ samples. Experimentally it was found that setting $\alpha = 0.5$ gave the best results in terms of IoU performance against the lidar labels. Unless otherwise stated, the model evidence importance was set to $\omega = 1$.

B. Detection Performance of Deep ISM vs Classical Filtering Methods

We compare the detection performance of our approach against cell averaging CFAR [11] applied in 1D (along range) for polar scans and in 2D for Cartesian scans by determining the quantity of occupied and unoccupied space successfully segmented in comparison to the ground truth labels generated from lidar in observed space. Due to class frequency imbalance, we use the mean Intersection Over Union (IoU) metric [22]. The optimum number of guard cells, grid cells and probability of false alarm, for each CFAR method, was determined through a grid search maximising the mean IoU of each approach on training data. For our method, each cell was judged as occupied or free based on a 0.5 probability threshold on $\boldsymbol{p}_{\boldsymbol{y}|\boldsymbol{x}}$. A 2m square in the centre of the occupancy map, corresponding to the location of the survey vehicle, was marked as unobserved.

The results form the test data set for each approach are shown in table I and show that our approach outperforms all the tested CFAR methods, increasing the performance in occupied space by 0.11, whilst achieving almost the same performance in free space leading to a mean IoU of 0.63. Our model is successfully able to reason about occupied space in light of challenging noise artefacts. In contrast, the challenge in free space is not in identification, with free space typically being characterised by low power returns, but in distinguishing between observed and occluded regions, a challenge which is missed entirely by the IoU metric. Figure 6a shows how our model is able to successfully determine space that is likely to be unknown because of occlusion and is able to clearly distinguish features, such as cars that are largely missed in CFAR. An occupancy grid of size 600×600 can be generated at around 14Hz on a NVIDIA Titan Xp GPU. Which is significantly faster than real time for radar with a frequency of 4Hz.

C. Uncertainty Prediction

As described in Section III-E, by incorporating aleatoric uncertainty into our formulation, the latent uncertainty associated with each grid cell is allowed to vary by predicting the standard deviation of each cell $\boldsymbol{\gamma}_\phi(\boldsymbol{x})$ alongside the predicted logit $\boldsymbol{\mu}_\phi(\boldsymbol{x})$. In this section we investigate the uncertainties that are captured by this mechanism.

To do this we gradually increase a threshold on the maximum allowable standard deviation of each cell $\boldsymbol{\gamma}_\phi(\boldsymbol{x})$ labelling any cell that falls below this threshold as either occupied (white) or free (black), whilst every cell above the threshold is labelled as unknown (grey). The result of this process is illustrated in Figure 6d.

The standard deviation predicted by our network largely captures uncertainty caused by occlusion, which, independent of the true underlying state of occupancy, results in space that is inherently unknown. From least likely to most likely to be occluded, we move from high power returns labelled as occupied, to a region nearby and up to the first return, to space that lies in partial and full occlusion. This ray tracing mechanism is largely captured by the standard deviation $\boldsymbol{\gamma}_\phi(\boldsymbol{x})$ predicted by our network.

D. Qualitative Results

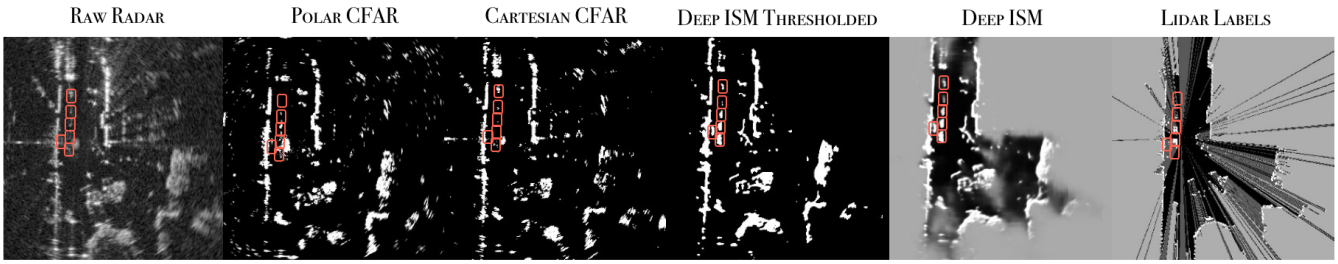
Finally, we provide several qualitative examples of our model operating in challenging real world environments and investigate how the strength of our prior term in (9) effects the occupancy distribution predicted by our model.

Figure 6c gives qualitative examples taken from the test set. Our network is able to successfully reason about the complex relationship between observed and unobserved space in light of challenging noise artefacts. In Figure 6b we vary the relative importance between the likelihood and KL divergence term by varying the hyper-parameter ω in (9). Increasing ω increases the relative importance of the likelihood term and leads to an ISM which is able to more freely reason about regions of space for which no labels exist during training, using the labels available in the observed scene. In the limit, of high ω the model is no longer able to successfully identify regions of space that are likely to be occluded, predicting all low power returns as free with a high probability.

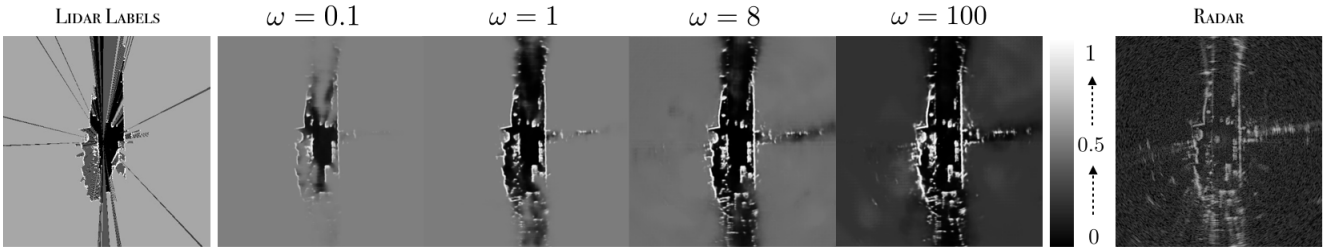
V. CONCLUSION

By using a deep network we are able to learn an inherently probabilistic ISM from raw radar data that is able to identify regions of space that are likely to be occupied in light of complex interactions between noise artefacts and occlusion. By accounting for scene context, our model is able to outperform CFAR filtering approaches. Additionally, by modelling heteroscedastic uncertainty we are able to capture the variation of uncertainty throughout the scene, which can be used to identify regions of space that are likely to be occluded. Our network is self-supervised using only partial labels generated from a lidar, allowing a robot to learn about the occupancy of the world by simply traversing an environment.

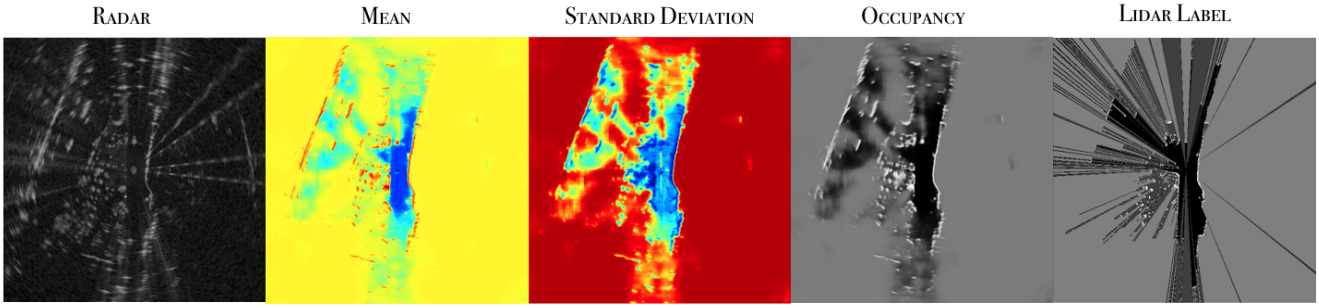
At present our approach operates under a static world assumption. In future work we hope to incorporate scene dynamics into our formulation allowing a robot to identify cells that are likely to be dynamic in addition to occupied or free.



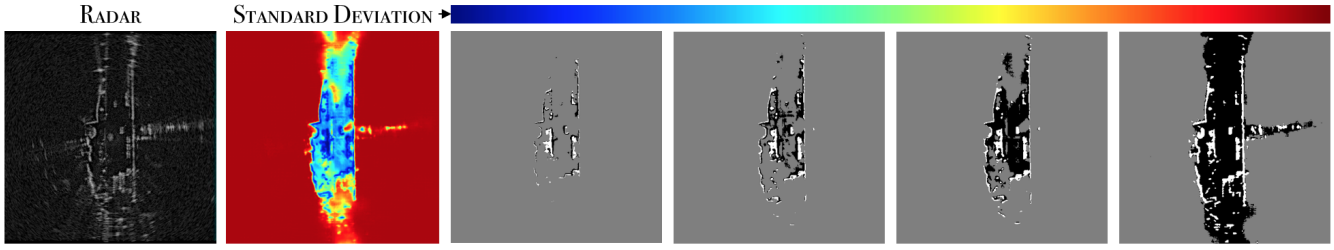
(a) The detection performance of our approach vs classical filtering methods with black representing predicted free and white representing predicted occupied by each approach. In comparison to CFAR our approach results in crisp and clean detections in observed and unobserved space. The red rectangles highlight cars that are clearly detected by our approach which are largely missed by CFAR. In addition, our model is able to successfully reason about what in the scene is likely to be unknown due to occlusion.



(b) The predicted probability of occupancy for different values of likelihood importance ω. As ω is increased our model becomes increasingly less conservative, reasoning in the unobserved region of space based on labels in the observed region.



(c) Our model successfully identifies occupied free and occluded space in challenging real world environments.



(d) A scene segmented as predicted occupied (white), unoccupied (black) and unknown (grey) for decreasing confidence thresholds (left to right) on the predicted standard deviation γ_ϕ. From most certain to most least certain, we move from high power returns labelled as occupied, to a region nearby and up to the first return, to space that lies in partial and full occlusion.

Fig. 6.

ACKNOWLEDGMENT

The authors would like to thank Oliver Bartlett and Jonathan Attias for proof reading a draft of the paper, and Dan Barnes for many insightful conversations, and the reviewers for helpful feedback.

This work was supported by training grant Programme Grant EP/M019918/1. We acknowledge use of Hartree Centre resources in this work. The STFC Hartree Centre is a research collaboratory in association with IBM providing High Performance Computing platforms funded by the UKs investment in e-Infrastructure. The Centre aims to develop and demonstrate next generation software, optimised to take advantage of the move towards exa-scale computing.

REFERENCES

[1] A. Elfes, "Using occupancy grids for mobile robot perception and navigation," *Computer*, no. 6, pp. 46–57, 1989.

[2] K. Konolige, "Improved occupancy grids for map building," *Autonomous Robots*, vol. 4, no. 4, pp. 351–367, 1997.

[3] A. Milstein, "Occupancy grid maps for localization and mapping," in *Motion Planning*, InTech, 2008.

[4] D. Filliat and J.-A. Meyer, "Map-based navigation in mobile robots:: I. a review of localization strategies," *Cognitive Systems Research*, vol. 4, no. 4, pp. 243–282, 2003.

[5] J.-A. Meyer and D. Filliat, "Map-based navigation in mobile robots:: Ii. a review of map-learning and path-planning strategies," *Cognitive Systems Research*, vol. 4, no. 4, pp. 283–317, 2003.

[6] B. Clarke, S. Worrall, G. Brooker, and E. Nebot, "Towards mapping of dynamic environments with fmcw radar," in *Intelligent Vehicles Symposium Workshops (IV Workshops), 2013 IEEE*, pp. 140–145, IEEE, 2013.

[7] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," in *Advances in Neural Information Processing Systems*, pp. 5580–5590, 2017.

[8] K. Werber, M. Rapp, J. Klappstein, M. Hahn, J. Dickmann, K. Dietmayer, and C. Waldschmidt, "Automotive radar gridmap representations," in *Microwaves for Intelligent Mobility (ICMIM), 2015 IEEE MTT-S International Conference on*, pp. 1–4, IEEE, 2015.

[9] R. Dia, J. Mottin, T. Rakotovao, D. Puschini, and S. Lesecq, "Evaluation of occupancy grid resolution through a novel approach for inverse sensor modeling," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 13841–13847, 2017.

[10] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autonomous robots*, vol. 15, no. 2, pp. 111–127, 2003.

[11] M. Skolnik, *Radar Handbook, Third Edition*. Electronics electrical engineering, McGraw-Hill Education, 2008.

[12] S. B. Thrun, "Exploration and model building in mobile robot domains," in *Neural Networks, 1993., IEEE International Conference on*, pp. 175–180, IEEE, 1993.

[13] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[15] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, pp. 3483–3491, 2015.

[16] H. R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "Deep learning and its application to medical image segmentation," *Medical Imaging Technology*, vol. 36, no. 2, pp. 63–71, 2018.

[17] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artificial Intelligence Review*, pp. 1–18, 2018.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[19] C. Esteves, C. Allen-Blanchette, X. Zhou, and K. Daniilidis, "Polar transformer networks," *arXiv preprint arXiv:1709.01889*, 2017.

[20] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.