

milliMap: Robust Indoor Mapping with Low-cost mmWave Radar

Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen,
Niki Trigoni, and Andrew Markham

Abstract—Single-chip Millimetre wave (mmWave) radar is emerging as an affordable, low-power range sensor in automotive and mobile applications. It can operate well in low visibility conditions, such as in the presence of smoke and debris, fitting the payloads of resource-constrained robotic platforms. Due to the nature of the sensor, however, distance measurements are very sparse and affected by multi-path reflections and scattering. Indoor grid mapping with mmWave radars has not been yet explored. To this extent we propose *milliMap*, a self-supervised architecture for creating dense occupancy grid maps of indoor environments from sparse, noisy mmWave measurements. To deal with the ill-constrained sparse-to-dense reconstruction problem, we leverage the Manhattan world structure typical of indoor environments to introduce an auxiliary loss that encourages generation of straight lines. With experiments in different indoor environments and under different conditions, we show the ability of *milliMap* to generalise to previously unseen environments. We also show how the reconstructed grid maps can be used in subsequent navigation tasks.

I. INTRODUCTION

The continued growth and evolution of mobile robotics applications demand increasing levels of autonomy and perception. In turn, advances in capability are also leading to the creation of novel human/robot systems, ranging from the niche (e.g. fire rescue) to the mundane (e.g. domestic service robots). For all these applications, navigation is a key capability and requirement.

State-of-the-art navigation and path planning approaches are often based on an occupancy map representation of the environment [5]. These maps are commonly built using laser range scanners (lidar), RGB-D cameras or stereo cameras. Although lidars provide high resolution point clouds, they are often impractical for low-cost, low-power applications. Camera-based sensors, on the other hand, whilst being relatively inexpensive, raise privacy concerns, particularly on consumer robotic platforms for domestic or commercial environments [23]. Meanwhile, use cases of vision sensors are also restricted by adverse illumination conditions, e.g., darkness, dimness and glare [6].

Recently, single-chip millimetre wave (mmWave) radar has emerged as an innovative low-cost, low-power sensor modality in the automotive industry. A key advantage of mmWave radar is its robustness to adverse environmental conditions, such as smoke, fog and dust. This unique capability makes it particularly useful in search and rescue scenarios, where teams of mobile robots operate in dark environments, full of airborne particulates. In the specific case of fire response, mmWave radars can see through smoke

All authors are with the Department of Computer Science, University of Oxford, UK. Emails: {firstname.lastname}@cs.ox.ac.uk

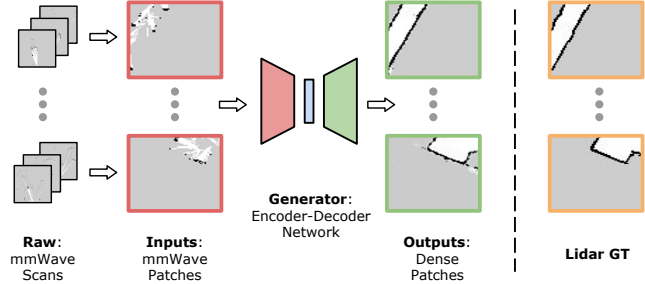


Fig. 1: An illustration of *milliMap*. A neural network generator takes as input the stitched patches from mmWave scans and produces denser and cleaner patches. Generated patches are then merged, yielding a dense grid map.

and help firefighters understand smoke-filled environments where many other sensors (e.g., RGB camera, depth camera and lidar) fail. Moreover, thanks to the use of beam-forming antennas rather than mechanical rotation, single-chip mmWave radar solutions are physically small and light. Compared with the cumbersome lidar or mechanical radar (e.g., CTS350-X), new mmWave radars are more able to fit the payloads of many micro robots and form factors of mobile or wearable devices.

Despite these advantages, mmWave-based mapping in indoor environments is still under-explored. The main issues lie in the strong indoor multi-path reflections as well as the sparse measurements returned by single chip radars. In extreme cases, we observe outliers due to multi-path reflections over 75%, along with an order of magnitude lower point density than a lidar counterpart.

To this extent, we propose *milliMap*, an approach for performing both denoising and a sparse-to-dense reconstruction of mmWave occupancy maps (see Fig. 1). Supervision labels can be provided as ground truth from a dense range sensor such as lidar. The system is then able to generalize to previously unseen environments. We show that, in order to learn an effective mapping, it is useful to introduce priors on the geometric appearance of indoor spaces, which are mostly composed of rectilinear features, such as walls and floors.

To summarize, the contributions of this work are:

- The first work using single-chip mmWave radars for dense grid mapping in indoor environments.
- A customized sparse-to-dense loss that embeds the geometric characteristics of indoor spaces.
- A systematic study of the impacts of input representations and network models on the map generation.
- Extensive experiments in various real-world settings, with dataset and code released to the community.


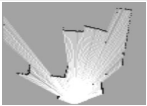

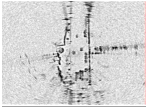


		Cost (\$)	Weight (kg)	Power (W)	Scan Points
	Lidar (VLP-16)	8,800	0.83	8	
	Mechanical Radar (CTS-350)	Customized Only	6	24	
	Our Radar (AWR1443)	40	<0.03	2	

Fig. 2: Comparison of lidar¹, mechanical radar² and our single-chip radar³. In each category, the features of a representative model are listed. Notably, compared with a lidar and a mechanical radar used in [30], our beamforming radar is much cheaper and lighter, but only provides few points.

II. RELATED WORK

RF Imaging and Tracking. Signal reflection of RF waves has been leveraged to perform object imaging in both WiFi and millimeter wave bands. In the WiFi bands, researchers have used commodity WiFi chips [3], [10] or specialized FMCW hardware [1], [32] to image static objects, or measure human body dynamics as well as pose estimation. However, due to the relatively narrow bandwidth and OFDM modulation, the performance of WiFi imaging methods is limited in the wild [10]. In contrast, because of the wide bandwidth in GHz and modulation specially designed for ranging rather than communications (e.g., FMCW), millimeter wave radars have been used for object imaging [2], [22]. However, these attempts all use a heavy mechanical radar in outdoor scenarios, where multi-path noise is insignificant. Imaging the indoor environments using single-chip mmWave radars is an important, yet unexplored area.

Sparse-to-Dense Generative Networks. Works that exploit sparsity have been proposed mainly for the problem of depth estimation from sparse depth measurements [9], [15], [17], [16], [29]. [9] exploits the sparsity of stereo disparity maps in the Wavelet domain. In [15], the authors leverage the regularities of indoor environments to infer dense depth from sparse measurements, based on compressive sensing. Other works focused on multi-modal inputs. [29] proposes to use a fully-connected conditional random fields model for depth inpainting from RGB and sparse (i.e., SLAM-derived or lidar measurements) or incomplete (Kinect-based) depth images. Completion of incomplete depth maps from structured light sensors is usually referred to as *depth inpainting*, or when the objective is also to remove measurement noise, *depth enhancement* [14]. [17] proposes a deep generative network for multi-modal depth prediction from RGB images and sparse depth measurements. In [16] the authors propose a self-supervised method for dense depth prediction from sequences of RGB and sparse depth images, based on photometric loss.

The most closely related work to our approach is [30], in

which the authors recently proposed a variational architecture for creating probabilistic occupancy grid maps from raw automotive scanning radar data. The difference lies in the distinct radars. Unlike [30], *milliMap* does not use a customized mechanical radar, but instead considers a cheap, lightweight beamforming radar commercial off-the-shelf (see Fig. 2). As stated in Section IV-A, the mmWave data in our case are much sparser. Moreover, [30] is designed for outdoor scenarios that do not suffer from the substantial multi-path present in indoor environments. We quantitatively compare our method with [30] in Tab. II.

III. PRINCIPLES OF MMWAVE RADAR

Range Measurement mmWave radar is based on the technique of frequency modulated continuous wave (FMCW) radar [26], and has the ability to simultaneously measure both the range and relative radial speed of the target. In FMCW, a radar uses a linear ‘chirp’ or swept frequency transmission. When receiving the signal reflected by an obstacle, the radar front-end computes the frequency difference between the transmitted reference signal and the received signal, which produces an Intermediate Frequency (IF) signal. Based on this IF signal, the distance d between the object and the radar can be calculated as:

$$d = \frac{f_{IF}c}{2S} \quad (1)$$

where c represents the light speed $3 \times 10^8 m/s$, f_{IF} is the frequency of the IF signal, and S is the frequency slope of the chirp. In the presence of multiple obstacles at different ranges, a fast Fourier transform (FFT) is performed on the IF signal, where each peak after FFT represents an obstacle at the corresponding distance.

Angle Measurement A mmWave radar estimates the obstacle angle by using *multiple* on-board antennas. It works by emitting chirps with the same initial phase, and then simultaneous sampling from multiple receiver antennas. Based on the differences in phase of the received signals, the Angle of Arrival (AoA) for the reflected signal can be estimated [20]. Formally, the AoA estimated from any two receiver antennas can be calculated as:

$$\theta = \sin^{-1}\left(\frac{\lambda\omega}{2\pi d}\right) \quad (2)$$

where ω denotes the phase difference and λ is the wave length. When multiple pairs of receiver antennas are available, the final AoA is the average result from different pairs. At this point, the position of a reflecting obstacle can be jointly determined by AoA and ranging estimation.

IV. PROPOSED APPROACH

In this section, we describe the technical details of *milliMap*. In Sec. IV-A, we introduce our technical challenges. The reconstruction methods, including the neural

¹<https://www.amtechs/product/VLP-16-Puck.pdf>

²<https://navtechradar.atlassian.net/wiki/spaces/PROD/pages/12353572/CTS350-X+Radar+Specifications>

³<http://www.ti.com/lit/ds/symlink/awr1443.pdf>

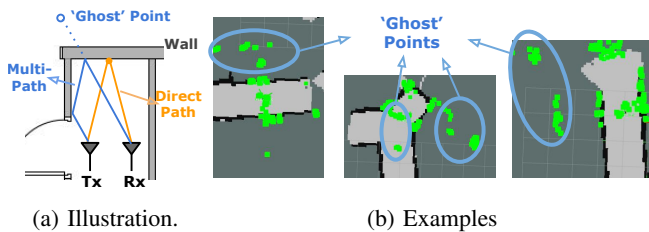


Fig. 3: Multi-path Noise. The black lines in (3b) are walls and there are non-negligible noise artefacts (in green) behind walls that are the result of multi-path reflection.

network architecture and learning fashion, are described in Sec. IV-B. Sec. IV-C discusses the network input representation and Sec. IV-D finally details a customized loss term and our objective function.

A. Challenges: Sparsity and Noise Issues

A mmWave radar detects ambient objects based on signal reflection. After several on-board pre-processing steps (e.g., interference mitigation), the range and orientation of reflecting points can be estimated and these points collectively form a *point cloud* in the field of view. However, unlike the dense point clouds generated by lidars or depth cameras, the mmWave point cloud in indoor environment has two fundamental issues: i) multi-path noise and ii) sparsity.

1) *Multi-path Noise*: Similar to any radio frequency technology, the signal propagation of mmWave in indoor environments is subject to multi-path [31] due to beam spreading and reflection from the surrounding objects (see Fig. 3a). As a consequence, reflected signals arriving at a receiver antenna are normally from two or more paths, leading to smearing and jitter. Multi-path is the primary contributor to the non-negligible proportion of pertinent noise artefacts or ‘ghost points’ in a mmWave point cloud. We empirically found that, in extremely severe multi-path scenarios, e.g., corridor corners, ghost points can account for $> 75\%$ points of a frame, which severely impacts grid mapping steps. Fig. 3b shows examples of noisy point clouds, where we can see many ghost points behind walls.

2) *Sparsity*: As shown in Fig. 2, the point cloud given by a single-chip mmWave radar is approximately ~ 100 reflective points per scan, which is over $10\times$ sparser than a lidar and $\sim 5\times$ sparser than a mechanical radar [16]. Such sparsity results from three factors in commercially available mmWave radars: (i) few antennas, (ii) point aggregation mechanism and (iii) restricted sensing range. Unlike massive array radar technology, due to cost and size constraints, the mmWave radar in our use only has 6 antennas, which fundamentally limits its resolution. In addition, unlike a mechanically rotating/scanning radar, the beamforming radar used in this work is static with limited field of view. Moreover, in order to lower bandwidth burden and improve signal-to-noise ratio, commercial mmWave radars usually apply algorithms such as CFAR (Constant False Alarm Rate) [28] on raw mmWave streams and *only* provide aggregated point cloud, further reducing density. The third factor resulting in sparsity is

specific to indoor mapping tasks and a consequence of multi-path noise. mmWave point clouds contain a non-negligible portion of ‘ghost points’, which can mislead map densification. In order to suppress these ‘ghost points’, we discard points outside of a sensing radius of 3m, as multi-path effects generally incur false-positive points at longer distances [31]. However, this restriction inevitably decreases the density of point clouds further.

B. Reconstruction Method

With knowledge of the properties of mmWave data, *milliMap* aims to combat the above issues and convert the raw mmWave point cloud to a grid map of occupancy. Although traditional Inverse Sensor Models (ISM) techniques work well on high-fidelity sensors such as lidar, these ISM methods struggle to model challenging radar noises and often impose strict assumptions on the noise distribution [30]. In fact, the complex interaction of noise and sparsity issues introduces huge challenges. As we will see soon in experiments, the map cannot be accurately reconstructed when the classic line-fitting approach [19] designed for lidar is used. In contrast, using deep learning methods, as originally advocated by [25], allows occupancy grids to be learned from raw data.

Reconstruction Neural Network. For these reasons, we adopt a deep learning method to reconstruct grid maps in this work. Our network architecture is constructed based on *pix2pixHD* [27], a proven encoder-decoder framework for general image-to-image translation. *pix2pixHD* is essentially a conditional generative adversarial network (GAN) [18] that comprises a generator G and a discriminator D . In our context, the goal of the generator G is to transform sparse and noisy patches to dense and clean images, while the discriminator D aims to distinguish real images (i.e., partial environment maps) from the transformed ones. As in many other generative networks, U-Net [21] is adopted as the backbone in our generator. To allow a large receptive field without large memory overhead, this network also uses multi-scale discriminators and downsamples the real and synthesized images by different factors to create an image pyramid of various scales. The discriminators are trained to distinguish real and generated images at various scales.

Self-Supervision by Co-location: Training the above neural network requires a large number of labelled images, which are costly to annotate by humans. To make *milliMap* scalable and reduce labelling effort, we adopt a self-supervised learning fashion by using only partial labels (i.e., lidar patches) generated from a co-located lidar, allowing a robot to learn about the occupancy of the indoor environment by simply traversing an environment. After the co-located learning phase, the mmWave radar on the robot is able to gain mapping skills from past experience and becomes capable of generating a lidar-like map independently. Fig. 4 illustrates this learning approach.

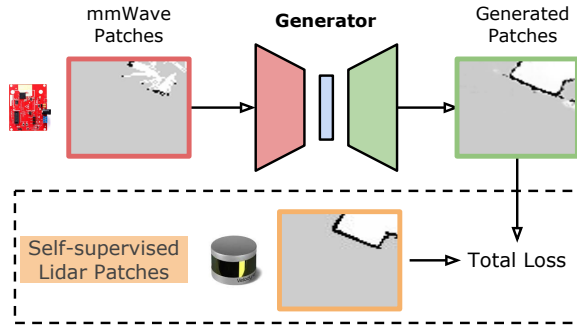


Fig. 4: Training the generator network with self-supervision.

C. Network Input Representation

Given the above neural network, it is not immediately clear what representation of the inputs is best suited. Similar to most networks for image-to-image translation, our network expects image-like inputs, with a fixed, relatively low, number of channels and spatial correlations between neighbouring pixels, which is not met by the inherent irregularity of point clouds. We thus need to firstly convert the point cloud to an image-like representation and then use existing networks to process it.

Perhaps the most straightforward representation is a virtual 2D laser *scan* obtained from the 3D point cloud. After projecting each scan to a planar 2D image via raytracing, generative convolution neural networks are able to take it as an input and generate a denser and denoised image. The dense images can then be converted back to angular distance measurements via raytracing and used for mapping. However, as the mmWave point cloud is very sparse, the converted scan image from each frame contains few spatial correlations between neighboring pixels. Directly feeding such non-informative images to a network often incurs overfitting and hard to generalize in new environments [24].

For these reasons, in this work we chose to work directly on map *patches*. In particular, we assume access to a reasonably accurate odometry (e.g., from fusion of wheel odometry and inertial measurements) and we directly generate a map from mmWave scans, using off-the-shelf Bayesian grid mapping. We then feed patches of the generated map along with the past robot trajectory to our network for denoising and densification. The advantage is that map patches contain more information about the structure of the environment; at the same time, mapping can be performed in real time, while the more expensive map densification process can run in background. Hereafter, we denote the real map patches as \mathbf{x} and the converted mmWave patches as \mathbf{s} . The pivotal goal of `milliMap` is to translate mmWave patches to real map patches through a deep neural network. Then given the generated dense patches, we stitch them together to produce a full grid map.

D. Objective Function

The objective function of our network comprises of losses from four sources: (1) conditional GAN, (2) intermediate

feature matching, (3) perceptual loss and (4) map prior. In particular, the *map-prior loss* is our proposed term that enforces indoor geometric consistency in the generated patches. **Reconstruction Likelihood.** We use conditional GANs to model the conditional distribution of real map patches \mathbf{x} given the input mmWave map patches \mathbf{s} , which are converted from the sparse point cloud. The conditional GAN loss can be expressed as:

$$\mathcal{L}_{cGAN}(G, D_k) = \mathbb{E}_{(\mathbf{s}, \mathbf{x})} [\log D_k(\mathbf{s}, \mathbf{x})] + \mathbb{E}_{\mathbf{s}} [\log(1 - D_k(\mathbf{s}, G(\mathbf{s})))]$$

where G tries to minimize this objective function against an adversary network D_k that tries to maximize it [18]. In particular, as our network uses multi-scale discriminators, D_k here is the specific discriminator for k -th scale. In the meantime, to stabilize training and generate meaningful statistics at multiple scales, we follow [4], [27] and introduce the feature matching loss $\mathcal{L}_{FM}(G, D_k)$ in our objective function:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(\mathbf{s}, \mathbf{x})} \sum_{i=1}^T \frac{1}{N_i} \|D_k^{(i)}(\mathbf{s}, \mathbf{x}) - D_k^{(i)}(\mathbf{s}, G(\mathbf{s}))\|_1$$

where T is the total number of layers, $D_k^{(i)}$ produces the features of i -th layer and N_i denotes the number of nodes in that layer. `milliMap` computes this feature matching loss on multiple discriminators which is in line with our multi-scale architecture. Lastly, to compare high level differences and stabilize GAN training [13], we also introduce a perceptual loss in the objective function:

$$\mathcal{L}_{VGG}(G) = \mathbb{E}_{(\mathbf{s}, \mathbf{x})} \sum_{j=1}^J \|F^{(j)}(G(\mathbf{s})) - F^{(j)}(\mathbf{x})\|_1$$

where F is a pre-trained loss network used for image classification that helps to quantify the perceptual differences of the content between images. In this work, we follow [13] and adopt the VGG network as F . Each layer j in the VGG network measures different levels of perception.

Map Prior. The above losses only consider the efficacy of reconstruction in the latent space of high-level appearance but ignore the important low-level geometrics. Recent research found that the latent spaces of appearance and geometry are not strongly correlated. Standard neural network generators can learn appearance transformation, however, lack the ability to embed complex geometry cues for effective image-to-image translation [8], [33]. Nevertheless, 2D indoor maps in modern buildings often have strong geometric structures that follow certain patterns, e.g. following rectilinear outlines for ease of construction. As this geometric information is fairly ubiquitous [7], one can leverage it as a prior to bootstrap the patch generation process and enhance the quality of the final stitched map. Formally, given a generated patch $G(\mathbf{s})$ and its corresponding real patch \mathbf{x} , we define a *map-prior loss* as follows:

$$\mathcal{L}_{MP}(G) = \mathbb{E}_{(\mathbf{s}, \mathbf{x})} \sum_{j=1}^M \|\mathbf{h}^{(j)} * G(\mathbf{s}) - \mathbf{h}^{(j)} * \mathbf{x}\|_1 \quad (3)$$

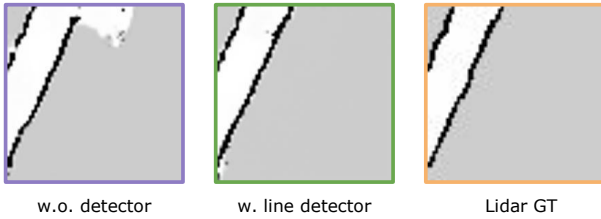


Fig. 5: Effectiveness of map prior loss on a straight corridor patch. A line detector is used in this case to construct the map-prior loss and the produced ‘corridor’ is straighter and more complete. lidar is used as pseudo-ground truth.

where $*$ represents the convolution operator and $\mathbf{h}^{(j)}$ is one of M convolution kernels with *fixed* weights, determined by the types of convolution. For example, $\mathbf{h}^{(j)}$ can be a line or edge detection mask, capturing different geometric properties of images. Through a detector mask, this map-prior loss encourages the consistency between source and target patches corresponding to a certain geometric prior. For example, many objects (e.g., walls and doors) on indoor floor plans are line based [7]. Therefore, when using line detectors to embed such a prior in the loss, we can achieve better reconstruction performances in corridors, as shown in Fig. 5. Choices of convolution masks are flexible, mainly depending on the noise level of inputs as well as a particular map/building type. We will discuss impacts of different types of detectors in Sec. VI-C.

Finally, our full objective combines reconstruction likelihood and map prior as:

$$\mathcal{L}_{total} = \sum_{k=1,2,\dots,K} \mathcal{L}_{cGAN}(G, D_k) + \lambda_1 \mathcal{L}_{FM}(G, D_k) + \lambda_2 \mathcal{L}_{VGG}(G) + \lambda_3 \mathcal{L}_{MP}(G) \quad (4)$$

where λ_1 , λ_2 and λ_3 are hyper-parameters for regularization. K denotes the number of distinct scales for discriminators.

V. IMPLEMENTATION

For the purpose of reproducing our approach, we release a novel dataset for indoor mapping with mmWave radars and the source code for `milliMap`⁴.

A. Dataset

A Turtlebot 2 platform endowed with multiple sensors is used as data collection platform. This dataset contains synchronized mmWave point cloud data from a TI AWR1443 board, lidar data from a Velodyne VLP-16 and wheel odometry. In addition, we provide RGB images from a front-facing monocular camera. The mmWave sensor, lidar and camera are coaxially located on the robot along the vertical axis. Two buildings are surveyed at the time of writing. The *Wolfson* building has a size of $\sim 1,100m^2$ and contains four floors, mostly composed of corridors and atriums; the *Robert Hooke* building (*RHB*) has a size of $\sim 150m^2$ and contains one floor with a combination of corridors and rooms. The *Wolfson*

TABLE I: Densification Before and After Mapping.

	Method	Wolfson		RHB	
		L1	IoU	L1	IoU
Scan (before)	Pix2Pix [11]	2.776	0.186	3.602	0.150
	Pix2PixHD [27]	2.309	0.226	2.722	0.152
Patch (after)	Pix2Pix [11]	2.214	0.319	3.200	0.173
	Pix2PixHD [27]	2.096	0.380	2.752	0.239

dataset presents a combination of walls, doors and large glass handrails; the *RHB* dataset presents walls, doors, glass panes and clutter. For each floor of the *Wolfson* building, we provide two runs along same corridors in opposite directions.

B. Training Details

Concerning network training, three loss weights λ_1 , λ_2 and λ_3 are set to 10, 10 and 5 respectively. We adopt a line detector as the convolution kernel in Eq. (3), M is set to 4, corresponding to 4 line directions in 0° , 45° , 90° and 135° . The training batch size is set to 16 and we use the Adam optimizer at a learning rate of $2e^{-3}$.

VI. EXPERIMENTAL EVALUATION

A. Evaluation Protocol

We now comprehensively evaluate `milliMap` through a set of experiments. Throughout this section, two metrics are consistently adopted to quantify map reconstruction performances: mean absolute error (L_1) and mean *intersection-over-union* (*IoU*), both of which are widely used [30]. We will omit ‘‘mean’’ hereafter for presentation ease. We perform cross-floor and cross-building tests to best examine the generalization ability and effectiveness of the trained model. Our data collection (see Sec. V-A) is divided into training and testing sets. In particular, the training set contains 12,000 augmented patch images extracted from maps of the 1st, 2nd and 3rd floors in *Wolfson* building. The data augmentation strategy we adopt here is the standard rotation and translation transformations on original patches to mitigate overfitting. Our test set comprises 49 patch images extracted from maps of the 4th floor in *Wolfson* building and 12 patches extracted from the 2nd floor of *Robert Hooke* building. All training and testing patch images have size 64×64 .

B. Impact of Densification Before and After Mapping

We first investigate the effect of two input representations (Section IV-C): (i) we perform densification of each scan and then aggregate them using grid mapping (denoted as *scan* representation) and (ii) we first aggregate scans using grid mapping and then perform densification on image patches (denoted as *patch* representation). As Tab. I shows, the reconstruction results of *patch* representation are significantly better than *scan* for both networks, implying the effectiveness of *patch* representation. Given the best-performing Pix2PixHD network, the L_1 errors of *scan* are 20% inferior to *patch*, with over 35% inferior IoU scores on both datasets. The reason is that the single *scan* densification easily overfits to straight lines, which is consistent to our discussion in Sec. IV-C.

⁴ <https://github.com/ChristopherLu/milliMap>

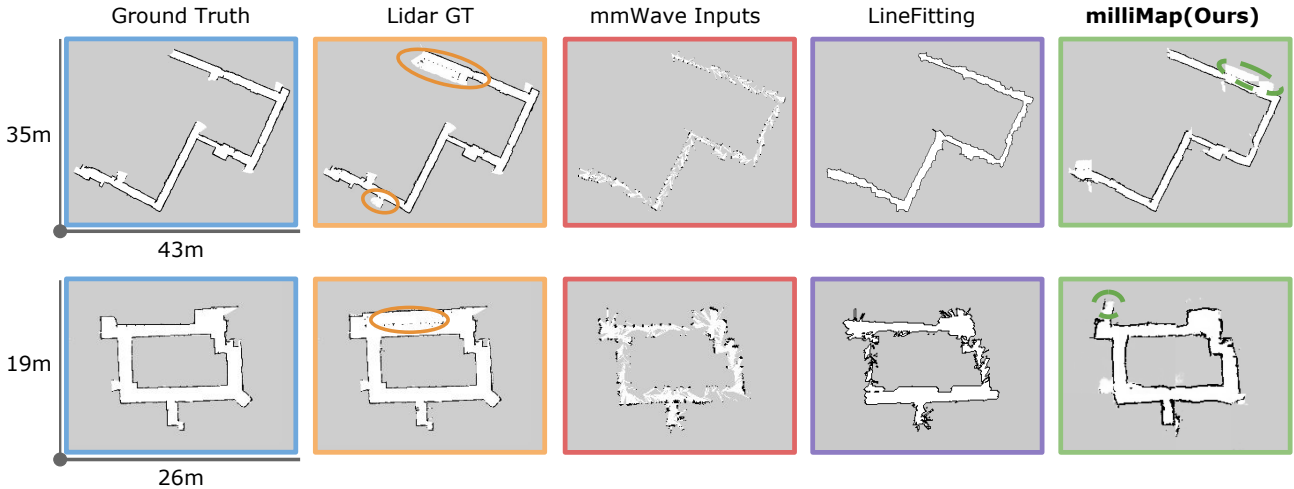


Fig. 6: Qualitative reconstruction results. *milliMap* achieves a comparable performance to the lidar counterpart. Solid circles on Lidar GT are glass objects; dashed circles are ‘ghost areas’ in generation. Top Row: *Wolfson*; Bottom Row: *RHB*.

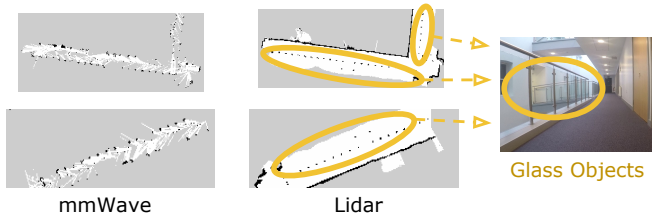


Fig. 7: Incorrect lidar supervision due to presence of glass objects in training data.

C. Network Architecture Validation

1) *Comparison*: After understanding the effective processing order, we adopt the *patch* representation for subsequent experiments and continue to validate different architectures of reconstruction networks. As *milliMap* is the first indoor mapping work dealing with very sparse inputs of such low-cost mmWave radar, we can only compare the following commonly used generative networks: Conditional Variational Autoencoder (CVAE) [30], BicycleGAN [34], Pix2Pix [11] and Pix2PixHD [27]. Notably, CVAE is the network architecture adopted by [30], though their goal is not sparse-to-dense due to the use of a customized mechanical radar. Beside these deep learning methods, we also compare with lineFitting [19], a classic reconstruction method for line-based indoor floor plans.

2) *Results*: Tab. II shows the performance comparison of different reconstruction methods. Despite its success on lidar map reconstruction, the classic line fitting method obviously struggles on both datasets and provides $< 50\%$ IoU than our approach, attributed to the substantial sparsity in raw mmWave maps. On the side of DNN methods, we did not find the advantages of using variational methods, implying that random sampling from a learnt distribution actually counteracts the benefits of uncertainty modelling and tends to output blurred reconstructions. We hypothesize that the performance gain can be also attributed to the strong regularity within indoor maps, which favours deterministic learning methods. Lastly, despite their close correlation, we

TABLE II: Reconstruction method comparison.

Method	<i>Wolfson</i>		<i>RHB</i>	
	L1	IoU	L1	IoU
LineFitting [19]	3.180	0.167	4.114	0.103
CVAE [30]	2.408	0.323	3.082	0.221
BicycleGAN [34]	2.538	0.303	3.393	0.195
Pix2Pix [11]	2.214	0.319	3.200	0.173
Pix2PixHD [27]	2.096	0.380	2.752	0.239
Ours	1.931	0.398	2.589	0.238

found that Pix2PixHD outperforms Pix2Pix on both datasets, thanks to the use of multi-scale discriminators and more losses. By introducing the map-prior loss, our method can further gain 9.6% L1 accuracy than Pix2PixHD, and better IoU performance overall on both datasets. Interestingly, in the last column of Fig. 6, there are ‘ghost’ areas on the generated maps, where part of a wall (black) is incorrectly marked as free regions (white). Recall that we adopt a self-supervision learning framework that uses lidar patches as supervision labels. These labels, however, can be error-prone when encountering glass objects (see the second column in Fig. 6), which is a commonly-known limitation of lidar. Although glass is opaque to mmWave, considering the high appearance similarity (see Fig. 7), we hypothesize the ‘ghost area’ of our generated *Wolfson* grid map can be attributed to the misleading lidar patches of glass in training. ‘Ghost’ areas do not appear with scan inputs, due to its overfitting to straight corridors.

D. Ablation Study

In order to examine the effect of different components in *milliMap*, we conduct an ablation study using different variants of our model. Our ablation study is dedicated to understand the impacts of two components: i) loss functions and ii) multi-scale discriminators.

1) *Loss Functions*: We modify the objective function of Eq. 4, by alternating different loss terms for reconstruction likelihood as well as alternating variants of our proposed map-prior term. Tab. III shows that the perceptual loss (i.e.,

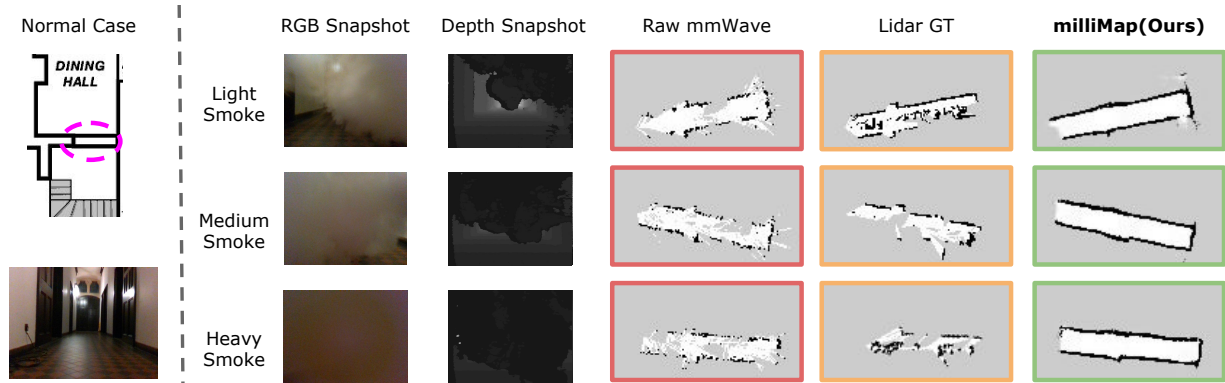


Fig. 8: Qualitative testing in smoke-filled environments.

TABLE III: Ablation study on losses and number of scales.

		<i>Wolfson</i>		<i>RHB</i>	
		L1	IoU	L1	IoU
Losses	w.o. FM	2.408	0.323	3.082	0.221
	w.o. VGG	2.538	0.303	3.393	0.195
	Edge Loss	2.214	0.319	3.200	0.173
# of Scales	1	2.024	0.394	2.633	0.250
	3	2.022	0.387	2.863	0.219
Ours		1.931	0.398	2.589	0.238

VGG loss) plays a vital role, and removing it incurs the largest performance decline ($\sim 30\%$) on both datasets. Feature matching loss is also necessary as it brings $16\% - 24\%$ gain in L_1 . These experiments indicate that, although grid maps are more about geometrics, these appearance losses are still important for stabilising generator training and improving realism. Interestingly, when we implement the map prior loss as edge detectors, its efficacy is not as helpful as the line detectors. This is because edges are a broad concept for any images and cannot effectively incorporate the geometrics of line-based maps. Moreover, as our supervision signals are from the imperfect lidar patches, the edge detectors are sensitive to the noises of lidar. In contrast, line detectors focus on low-frequency components of images and thus can be more robust to noise.

2) *Number of Scales*: Next we examine the impact of multi-scale discriminators. Recall that `milliMap` uses a 2-scale discriminator while our ablation study further examines the cases of 1- and 3-scales. As shown in Tab. III, the overall impact of multi-scale discriminators is not substantial ($\sim 5\%$) when varying the number of scales. This is as expected because the multi-scale discriminators were originally designed for high-resolution images while our input patches are not. We observed a marginal improvement from single-scale to 2-scale discriminators as more diverse feature matching is introduced in different scales. However, such increase of scales soon counteracts the benefits when the 3-scale network becomes oversized and overfits. This overfitting issue is more obvious on *RHB* dataset due to cross-building testing.

E. Mapping in challenging conditions

We now move on to the robustness analysis of map reconstruction by examining two challenging scenarios in

real world: (i) smoke-filled scenarios and (ii) noisy odometry.

1) *Smoke-filled Scenarios*: In this experiment we examine the potential use of `milliMap` in fire-fighting situations where other sensors fail (e.g., RGB cameras, depth cameras and lidars) due to smoke. To this end, we use a smoke machine to create different smoke densities in a corridor ($12 \times 1.5\text{m}^2$). Various sensor data were collected in both buildings, including lidar, depth cameras and mmWave radar. Fig. 8 shows the reconstructed map in 3 different smoke-filled scenarios. As we can see, lidar gives very inaccurate map results even with low levels of smoke. Due to the occlusion and reflection effects of smoke particles, lidar generates many non-existent obstacles and/or misses a lot of real ones. Depth cameras also face the same problem. In contrast, the mmWave radar is able to see through smoke and `milliMap` reconstructs the corridor accurately in all 3 smoke-filled scenarios. These results confirm the robustness of `milliMap` and we believe there are many promising use cases of it in search-and-rescue situations.

2) *Noisy-Odometry Scenario*: In this experiment, our goal is to test `milliMap`'s potential on hand-held devices, e.g., smartphones and tablets. Note that, for hand-held devices, their odometry is usually inferred from embedded microelectromechanical-inertial measurement unit by pedestrian dead reckoning (PDR) methods [12]. However, compared to wheel odometry, PDR odometry drifts more and has a lower sampling rate due to step discretization. As a consequence, the raw patch images of PDR are of lower fidelity. Furthermore, due to different viewpoints (e.g., different heights of robots and pedestrians), the mmWave observations have obvious differences from the training samples. Despite many compromising factors, as we can see in Fig. 9, `milliMap` still provides a reasonable reconstruction. Although such prediction is not accurate enough for robot navigation, it could potentially support some use cases for augmented reality on hand-held devices.

F. Downstream Navigation Tasks

We now test whether the produced maps, despite their imperfections, can still be used for autonomous navigation. In particular, we investigate if a robot is able to localize in the predicted map with comparable accuracy to that of a lidar

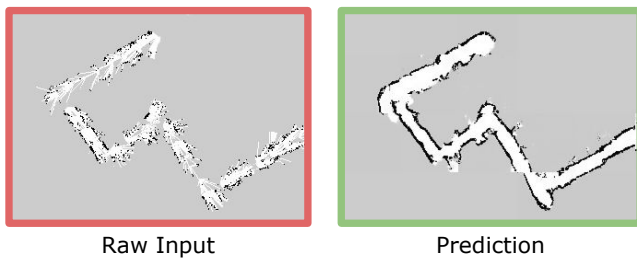


Fig. 9: Qualitative result for hand-held cases.

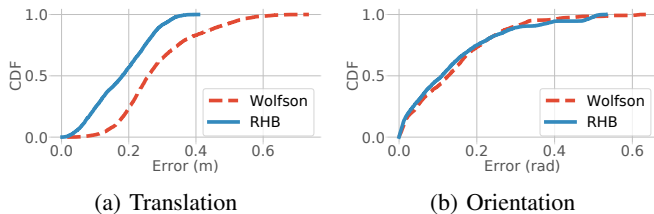


Fig. 10: Error CDFs for the downstream localization tasks.

map. We run Monte Carlo localization using mmWave raw measurements on the aforementioned reconstructed maps using the standard *amcl* ROS package with default parameters. Each time the robot starts at a random location. The pseudo-ground truth is derived by localization with lidar on a lidar map of the same floor. Fig. 10 shows the cumulative error distribution for 50 Monte Carlo runs. For the reconstructed *Wolfson* map, our robot achieved a mean translation accuracy of 0.285m and orientation accuracy of 0.142 rad; on the reconstructed *RHB* map, the mean translation and orientation accuracy are 0.178m and 0.140 rad respectively. Given the size of the two buildings, these results show that the map produced by *milliMap* can be used for higher-level tasks with excellent performance.

VII. CONCLUSIONS

We presented *milliMap*, a learning-based inductive method for obtaining dense occupancy grid maps from low-cost mmWave radar sensors, using self-supervision from partial labels from a lidar. By leveraging the structure of indoor scenarios, the model is able to reconstruct the shape of novel environments and, to some extent, cope with noisy odometry and smoke-filled scenarios. The limitation of the approach lies in the potential inaccuracy of labels (e.g., in presence of glass and reflective materials for lidar). Future work will be devoted to automatically detect such materials from the raw mmWave measurements, that are robust to presence of glass and metal.

REFERENCES

- [1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics*, 34(6):219, 2015.
- [2] Graham Brooker, David Johnson, James Underwood, Javier Martinez, and Lu Xuan. Using the polarization of millimeter-wave radar as a navigation aid. *Journal of Field Robotics*, 32(1):3–19, 2015.
- [3] Saandeep Depatla, Lucas Buckland, and Yasamin Mostofi. X-ray vision with only wifi power measurements using rytoV wave models. *IEEE Transactions on Vehicular Technology*, 64(4):1376–1387, 2015.

- [4] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 2016.
- [5] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.
- [6] Catherine Enright. Visual slam and localization—the hard cases. In *Electronic Imaging*, 2018.
- [7] Andrea Garulli, Antonio Giannitrapani, Andrea Rossi, and Antonio Vicino. Mobile robot slam for line-based environment representation. In *CDC*, 2005.
- [8] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving shape deformation in unsupervised image-to-image translation. In *ECCV*, 2018.
- [9] Simon Hawe, Martin Kleinsteuber, and Klaus Diepold. Dense disparity maps from sparse disparity measurements. In *ICCV*, 2011.
- [10] Donny Huang, Rajalakshmi Nandakumar, and Shyamnath Gollakota. Feasibility and limits of wi-fi imaging. In *SenSys*, 2014.
- [11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.
- [12] Antonio R Jimenez, Fernando Seco, Carlos Prieto, and Jorge Guevara. A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu. In *WISP*, 2009.
- [13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [14] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, 2014.
- [15] Fangchang Ma, Luca Carlone, Ulas Ayaz, and Sertac Karaman. Sparse sensing for resource-constrained depth reconstruction. In *IROS*, 2016.
- [16] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *ICRA*, 2019.
- [17] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *ICRA*, 2018.
- [18] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. In *arXiv preprint arXiv:1411.1784*, 2014.
- [19] Samuel T Pfister, Stergios I Roumeliotis, and Joel W Burdick. Weighted line fitting algorithms for mobile robot map building and efficient data representation. In *ICRA*, 2003.
- [20] Peng Rong and Mihail L Sichiitu. Angle of arrival localization for wireless sensor networks. In *SECON*, 2006.
- [21] Olaf Ronneberger, Philipp Fischer, and et al. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.
- [22] Raphaël Rouveure, Patrice Faure, and Marie-Odile Monod. Pelican: Panoramic millimeter-wave radar for perception in mobile robotics application. *Robotics and Autonomous Systems*, 81:1–16, 2016.
- [23] J. Stanley. The dawn of robot surveillance, June 2019.
- [24] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *ICLR*, 2015.
- [25] Sebastian B Thrun. Exploration and model building in mobile robot domains. In *ICNN*, 1993.
- [26] Deepak Uttam and B Culshaw. Precision time domain reflectometry in optical fiber systems using a frequency modulated continuous wave ranging technique. *Journal of Lightwave Technology*, 1985.
- [27] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.
- [28] DK Barton HR Ward. *Handbook of radar measurement*. 1969.
- [29] Chamara Saroj Weerasekera and et al. Just-in-time reconstruction: inpainting sparse maps using single view depth predictors as priors. In *ICRA*, 2018.
- [30] Rob Weston, Sarah Cen, Paul Newman, and Ingmar Posner. Probably unknown: Deep inverse sensor modelling in radar. In *ICRA*, 2018.
- [31] Yan Yan, Long Li, Guodong Xie, Changjing Bao, Peicheng Liao, Hao Huang, Yongxiong Ren, Nisar Ahmed, Zhe Wang, et al. Multipath effects in millimetre-wave wireless communication using orbital angular momentum multiplexing. *Scientific reports*, 6:33482, 2016.
- [32] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, and et al. Rf-based 3d skeletons. In *SIGCOMM*, 2018.
- [33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.
- [34] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, 2017.