RSS-Net: Weakly-Supervised Multi-Class Semantic Segmentation with FMCW Radar

Prannay Kaul, Daniele De Martini, Matthew Gadd, Paul Newman
Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK.
{prannay,daniele,mattgadd,pnewman}@robots.ox.ac.uk

***Abstract*—This paper presents an efficient annotation procedure and an application thereof to end-to-end, rich semantic segmentation of the sensed environment using Frequency-Modulated Continuous-Wave scanning radar. We advocate radar over the traditional sensors used for this task as it operates at longer ranges and is substantially more robust to adverse weather and illumination conditions. We avoid laborious manual labelling by exploiting the largest radar-focused urban autonomy dataset collected to date, correlating radar scans with RGB cameras and LiDAR sensors, for which semantic segmentation is an already consolidated procedure. The training procedure leverages a state-of-the-art natural image segmentation system which is publicly available and as such, in contrast to previous approaches, allows for the production of copious labels for the radar stream by incorporating four camera and two LiDAR streams. Additionally, the losses are computed taking into account labels to the radar sensor horizon by accumulating LiDAR returns along a pose-chain ahead and behind of the current vehicle position. Finally, we present the network with multi-channel radar scan inputs in order to deal with ephemeral and dynamic scene objects.**

***Index Terms*—perception, radar, semantic segmentation, deep learning, weakly-supervised learning**

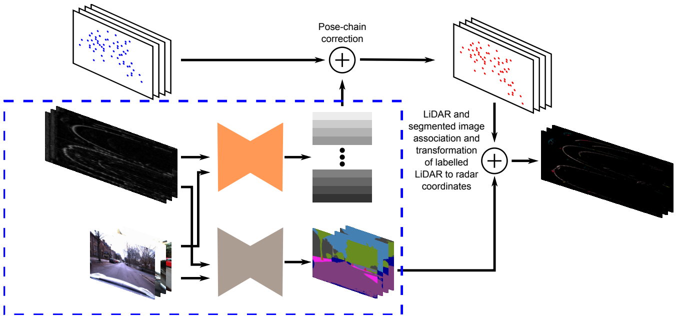

Pose-chain correction
LiDAR and segmented image association and transformation of labelled LiDAR to radar coordinates

Figure 1. An overview of the pipeline implemented to generate labelled training data for radar segmentation. The section within the blue box is completed before training such that segmented RGB streams are available on disk, as is the pose chain described in Section III-C. During training, a radar scan is selected from the training set. The temporally nearest RGB images and corresponding LiDAR scans are then used to form the labelled radar image as described in Section III. The resulting data are therefore formed on the fly during the training/testing process.

I. Introduction

Safe navigation and operation of mobile robots in search and rescue, agriculture, and mining environments will require perception systems that deliver a detailed understanding of the surroundings regardless of adverse environmental factors.

Light Detection and Ranging (LiDAR) and vision-based systems have been widely investigated and adopted in the last decade. However, these models are usually trained on datasets such as [1] which are captured in uniform conditions, leaving the state-of-the-art susceptible to rain, snow, fog, glare, lighting, and seasonal appearance changes.

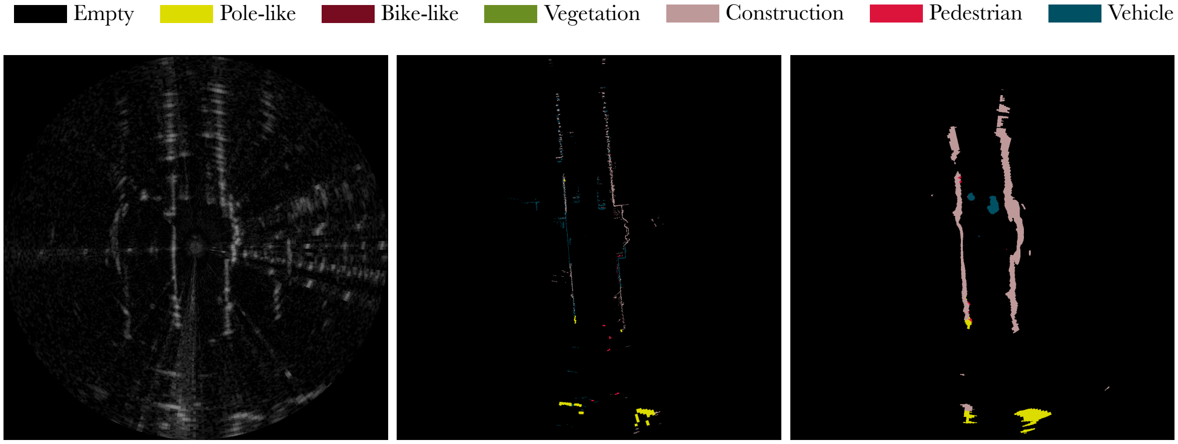
In contrast, due to the long wavelength of radio waves, Frequency-Modulated Continuous-Wave (FMCW) scanning radar operates well under variable weather and lighting conditions. Additionally, multiple returns are received from a single azimuthal transmission and operation at ranges of up to hundreds of metres is common. Indeed, there is a burgeoning interest in exploiting FMCW radar to enable robust mobile autonomy, including ego-motion estimation [2]–[6] and localisation [7], [8]. However, the radar measurement process is complex and scans formed thereby are prone to pollution by multipath reflections, speckle noise, and other artefacts in addition to the internal noise characteristics [9]. Furthermore, radar scans compress the 3D environment into a 2D planar representation, which limits the discrimination between objects which have similar appearance.

Figure 2 shows an urban scene as perceived by a FMCW radar and the corresponding segmentation from the proposed system. Section IV presents a method to segment this raw radar scan using a Fully Convolutional Neural Network (FCNN). This system is designed with the principle that sensor artefacts, which may appear similar to actual objects, are to be ignored and objects with distinct typical dynamics should be reliably and unambiguously identified. As even radar experts find labelling non-intuitive, Section III proposes a novel annotation technique suitable for weak supervision using alternative sensing modalities also present on the mobile robot (RGB and LiDAR streams), leveraging a state-of-the-art, publicly available, pre-trained model.

II. Related Work

The success of Deep Learning (DL) in general perception tasks, such as object classification and semantic segmentation of natural images, is well documented in the community [10], benefits of which are now impacting radar sensing [6]–[8].

Occupancy grid methods [11] provide discrimination between free and occupied space, dealing with the complex sensor artefacts encountered in radar scans, but fail to deliver richer information regarding the nature of the occupied space. To overcome this, a FCNN is used in [12] to segment a radar scan based on the probability of occupancy and hand labels are used to classify objects. However, hand labelling of



(b) (c) (d)

Figure 2. Comparison of a (b) raw radar scan, (c) labelled targets for training the network, and (d) predicted semantic segmentation, all shown in Cartesian representation. Here, the Cartesian representation takes the form of an image with resolution $N \times N = 1000 \times 1000$, where N is the cardinality of the set of bins in the discretised range returns of the original polar representation.

radar scans is a time-consuming task; therefore, as a trade-off between quantity and quality of labelled data, we use a form of weak supervision: semantically segmenting camera streams and combining the result with LiDAR range measurements to provide a labelled image in the radar frame.

Feature-based clustering approaches have been shown to work in [13], while in [14] an ensemble, where each classifier takes as input its own specialised feature set, allows for unseen patterns to be detected. In contrast to [13], [14] and similarly to [15] our work does not use hand-crafted features nor ensembles of classifiers. Furthermore, in contrast to [15] we do not use deep networks designed for pointclouds and instead use popular image segmentation networks on the the radar scans. Moreover, in contrast to all of [13]–[15] our work does not require manual labelling.

From the definitions in [16], our method falls in both *inexact* – as the radar data and LiDAR data have different granularities and range – and *inaccurate* – as the labels are not created by a human supervisor.

An example of both inexact and inaccurate supervision applied to radar is [11], where the authors use a U-Net, the popular image segmentation network [17], to segment a radar scan based on probability of occupancy on a more extensive dataset and produces good results. In contrast to [11], however, where the shorter maximum range of the LiDAR is not dealt with, we accumulate LiDAR returns in the native radar representation using a good external source of ego-motion.

A second example of inaccurate supervision can be found in [4]: an automatic labelling procedure is carried on to classify readings in radar scans as consistent across wide baselines or not; a U-Net is then trained to predict pixels in the polar representation of the radar scan. The result is a filtering technique, which limits the number of landmarks for the odometry pipeline, making the pipeline faster without losing accuracy.

Finally, in the same spirit as us, the authors of [18] apply a Weakly-Supervised Learning (WSL) approach to train a network for object detection on dynamic grid maps. They exploit temporal and spatial relationships to extract moving objects and their shapes using a LiDAR sensor. Indeed, once an object is observed, it continues being observed until it exits the field of view, updating shape and trajectory information. Then, this information is propagated backwards in time to refine the annotation for more consistency in the labels.

III. Training Data Generation

Figure 1 provides an overview of the pipeline used to generate training data for each of the radar scans. We leverage the depth of research by the computer vision community in the semantic segmentation of RGB images and specifically of urban street scenes. The offline datasets we use contain concurrent RGB image streams covering the full azimuth range (four cameras with 360° horizontal field-of-view), LiDAR scans from two lasers, and radar scans.

A. Semantic Image Segmentation

In order to create labelled images for the training procedure, we use the publicly available Deeplabv3-DPC [19] model, trained on the Cityscapes dataset, to perform semantic image segmentation on monocular visual image streams, identifying each pixel as belonging to one of a number of meaningful object categories. The Cityscapes dataset is comprised of forward facing urban street scenes, in contrast to our RGB streams which contain camera streams in four directions (forward, rear, left, and right). Furthermore, each stream is captured during dynamic motion and so many images are impacted by motion blur (particularly near the edges) and exposure effects.

B. LiDAR and RGB Image Fusion

Given the rich semantic segmentation extracted from the RGB streams, a method is required to use this information to generate the labelled radar training data. Using extrinsic transformations between the seven sensors (four RGB cameras, two

LiDAR scanners and one radar) and the intrinsic parameters of the four RGB cameras, one can project the LiDAR pointcloud at a given time-step onto each corresponding camera image. By representing the pointcloud in image coordinates a label for each point within the image can be extracted using the segmented images. After associating a label with each point of the LiDAR scan, using each of the four camera streams, the extrinsic transformation between the relevant LiDAR scanner and the radar is used to form labelled data in the radar coordinate frame. Finally these points are projected into the horizontal plane and discretized into range-azimuth grid cells. For the case in which multiple labelled points map to the same pixel location, the label is selected with equal probability.

C. *Pose-chain Interpolation of Labels*

The LiDAR scans from each sensor are gathered at $20\,\mathrm{Hz}$, whereas the camera and radar streams are gathered at $25\,\mathrm{Hz}$ and $4\,\mathrm{Hz}$, respectively. Furthermore there is a non-negligible difference between the times at which the radar scan is taken and the temporally closest four images (one image for each direction) and LiDAR scan. Due to this temporal difference and the dynamic nature of the environment which the sensors are operating within, the projection of the pointcloud into a given camera image suffers from misalignment.

We correct for this misalignment through interpolating an optimised pose chain between the time of LiDAR capture and its temporally closest images, one from each camera. Given this pose-chain, the required transformation is obtained through interpolation. The rotational component is obtained by Spherical Linear Interpolation (SLERP) on a spherical surface traced by a unit quaternion, as described in more detail in [20], and the translational component is obtained by a constant velocity interpolation.

In the same fashion, each radar scan is related to the closest LiDAR pointclouds in time. Indeed, for each azimuth message, a pointcloud is selected from each sensor ad projected into the scan though motion interpolation and extrinsic transformations.

D. *Accumulation of labels to the radar sensor horizon*

As seen in Figure 2(c), LiDAR scans are an inherently sparse representation of an environment. Furthermore, as the datasets used are collected in urban environments, the result is the majority of pointcloud readings are within a relatively short range compared to the full operating range of a FMCW radar and very few meaningful labels available at a distance past $40\,\mathrm{m}$. Already having access to the pose-chain described in Section III-C, we employ it once again to transform labelled LiDAR pointclouds before and after the current radar scan along the trajectory of the robot into the radar frame.

IV. Network Architecture for Radar Semantic Segmentation

Figure 4 provides an illustration of the FCNN architecture we employ. The design makes use of dilated convolutions initially, in the encoder section (orange channels) to increase the receptive field exponentially. The sparse nature of the generated training data requires fine details to be managed by the network, therefore only 4 max-pooling layers are used in total to prevent a high loss of detail from the input image, making the smallest feature map have dimensions sixteen times smaller than the input. The rich features after four sections of encoding are passed through an atrous (dilated) spatial pyramid pooling (ASPP) block introduced in [21], which uses varying rates of dilated convolution and global average pooling to produce rich semantic information at various scales. The output feature maps of these different dilated convolutions are concatenated together and reduced to 256 channels using 1×1 convolutions. The bilinear upsampled versions of these feature maps are concatenated with higher resolution feature maps from earlier in the network. This provides a mix between rich semantic information and fine detailed feature maps. Subsequent convolutions are used to reduce 304 feature maps to L channels which are bilinearly upsampled to give a final output with size equal to the input. During training, the L output channels are passed into a cross entropy loss function along with the true labels (generated as described in Section III). During testing, the L output channels are passed through an ‘argmax’ layer yielding a single channel output with class labels assigned to each element/pixel. In the above explanation, L is the number of classes segmented by the network.

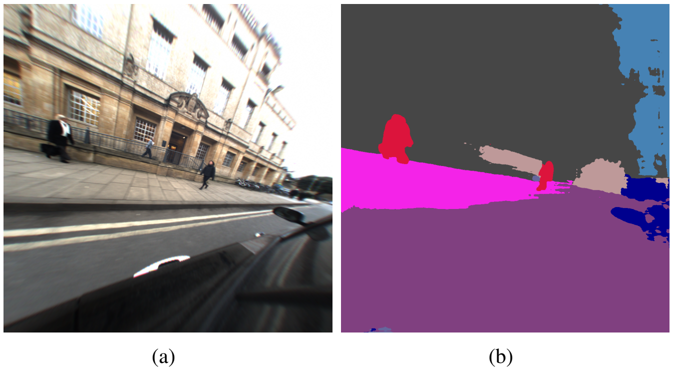

(a) (b)

Figure 3. Example of *one of* the four RGB streams, extracted at the same time-step, with the corresponding segmentation extracted from Deeplabv3-DPC. In this example, taken from one of the wide-angle monocular cameras, one observe motion blur, non-ideal exposure and a reduction in segmentation quality as the edge of the image is approached.

A. *Class Weighting*

As the target images are generated from LiDAR scans which are inherently sparse, there is a large class imbalance in the dataset. To overcome this, the loss function is weighted depending on the true class using a logarithmic function:

$$w[i] \propto \left(1 + \log \frac{\sum_{j=1}^{N} t[j]}{N \cdot t[i]}\right)^2 \tag{1}$$

where $w[i]$ is the class weight, $t[i]$ is the number of pixels belonging to class i in the training set and N is the number of classes. It should be noted that the weight for Empty class (see Section V-D) is empirically set to 0.1.

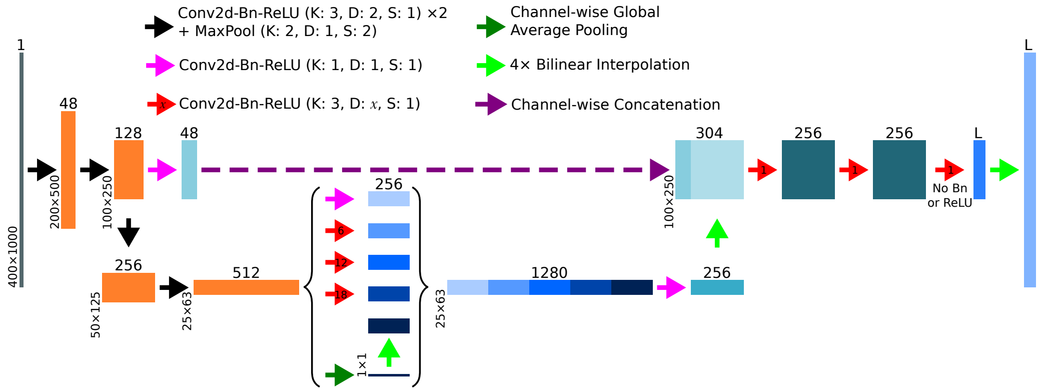

Conv2d-Bn-ReLU (K: 3, D: 2, S: 1) ×2
+ MaxPool (K: 2, D: 1, S: 2)
Conv2d-Bn-ReLU (K: 1, D: 1, S: 1)
x
Conv2d-Bn-ReLU (K: 3, D: x, S: 1)
Channel-wise Global
Average Pooling
4× Bilinear Interpolation
Channel-wise Concatenation
1
400×1000
48
200×500
128
100×250
48
256
50×125
25×63
512
256
6
12
18
1×1
25×63
1280
256
304
100×250
256
256
1
1
1
L
No Bn
or ReLU
L

Figure 4. Overview of the network used for radar semantic segmentation. L–number of classes. Bn–2D batch normalization, ReLU–Rectified Linear Unit, K–kernel size (all kernels are square), D–dilation rate, S–convolution stride.

B. Addressing Dynamic Objects

Different classes in a single radar scan often appear near-identical (e.g. a "Pedestrian" and "Pole-like") even to an expert, they exhibit spatial similarity. The discriminating feature in such cases is the dynamics of the object – a "Pole" is stationary, but a "Pedestrian" is (usually) moving – they exhibit temporal dissimilarity. To improve performance in such cases, three consecutive radar scans are input to the FCNN, providing the network with temporal information. Pose-chain interpolation is used to transform the scans into the same frame of reference; this ensures that stationary objects appear in the same location across all three scans. Given the pose-chain is in Cartesian coordinates, the radar scans are input to the network in Cartesian form.

Despite adding a temporal component the network remains fully convolutional and no recurrent units are used. To aid the network to learn useful distinct temporal and spatial kernel weights, depthwise separable convolutions are used throughout the network. Depthwise separable convolutions provide the added benefit of reduced network parameters and increased training time [22].

V. Experimental Setup

The experiments are performed using data collected from the *Oxford RobotCar* platform [23].

A. Ground truth radar ego-motion

We use the ground truth pose dataset described in the recently released *Oxford Radar RobotCar Dataset* [24] which is computed by an optimisation using Global Positioning System (GPS), robust Visual Odometry (VO) [25], and visual loop closures from FAB-MAP [26].

B. Dataset Demarcation for Training and Validation

The route in Oxford city centre used for collecting the offline datasets is taken from [8]. The path has been divided into three different portions: *train* (blue), *validation* (black) and *test* (purple). We specifically designed the sets in such a way that no intersection would occur among the three of them. Although the radar scanner can sense up to hundreds of metres, the environment taken in consideration is cluttered and does not allow it to reach such sensing ranges; thus we decided not to discard any sample on the dataset, but only to add a little padding between the portions, of the order of ten metres. The resulting dataset is then formed by 6246 training, 314 validation and 1720 test examples.

C. Data Augmentation

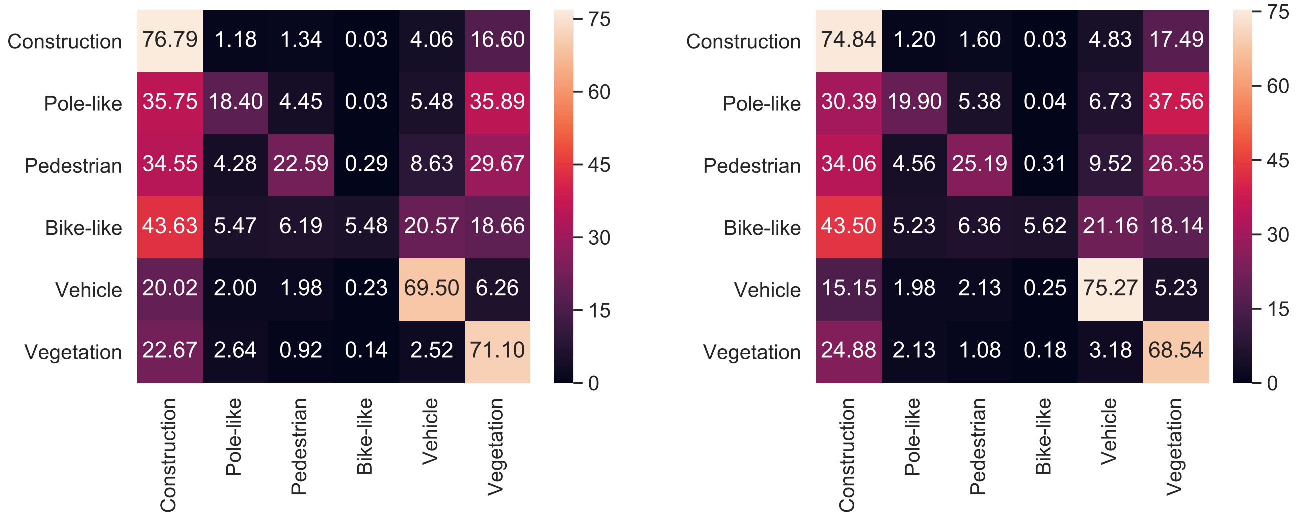
Data augmentation has been taken into account in the training phase to increase the number of training examples. We simply add a random flips on the horizontal and vertical axis with a 50% probability.

D. Class Definitions

The Cityscapes dataset uses 34 distinct classes with 19 recommended for use during training. Use of such a rich class dictionary is not suitable for radar scans. For example, the Cityscapes dataset makes distinction between buildings, walls and fences. The difference between these classes are rarely based on dimension, but instead on appearance and so in a radar scan they appear nearly identical. A similar case can be made for traffic signs, street lights and poles. Furthermore, certain classes are unobservable in radar scans, such as roads, sky and grass areas. For these reasons certain classes are omitted when generating the labelled training data and others are grouped together. Although the road is not observable, it can be inferred to from the robot location, which in turn could provide more information regarding the presence of sidewalks or grass terrain. Including such priors is not addressed in this paper and is left for future work.

In total the network uses 7 object classes. These are grouped in the following way with the labels from the Cityscapes dataset shown in parentheses:

1) Empty (sky, road, sidewalk, terrain, guard rail etc.)
2) Construction (building, wall, fence)
3) Pole-like (poles, traffic lights, traffic sign)
4) Pedestrian (person)
5) Vehicle (car, truck, bus, caravan, trailer, train)
6) Bike-like (rider, bicycle, motorcycle)
7) Vegetation (vegetation)



(a)

(b)

Figure 5. Normalized confusion matrix on the test set considering (a) the full radar sensing horizon and (b) a foreshortened radar sensing horizon.

Although “Pole-like” objects and “Pedestrians” generally appear similar in radar scans, we decided to maintain separate classes, since different vehicle behaviour is needed when either poles or people are present.

VI. Results

Figure 5(a) shows the distributions of predicted and true class memberships. Most classification errors arise from non-Construction objects being labelled as Construction objects. Moreover, Figure 5(b) shows the confusion matrix when only considering the closest parts of the environment. The relatively small difference in performance demonstrates that the network has generated representations which are equivariant in the polar form of the radar scans and understands the varying appearance of objects with changing radial distance from the robot.

Figure 2 shows an example from the test set (described in Section V-B). One can see from the middle row that a large part clearly in the class ‘Construction’ has been incorrectly labelled as the Vehicle class (observe the bottom left portion of the polar image). This arise due to two reasons; the first is that despite improving the projection of the LiDAR pointcloud into the image streams, using a pose-chain there will still be a discrepancy between the alignment of the image and the pointcloud. This leads to incorrect labelling of some regions, particularly at the boundaries between objects where the depth of the scene changes most rapidly. Secondly, the segmentation of the RGB images itself is not perfect and so even if the alignment between images and pointclouds were exact, incorrect or bloated segmentation labels lead to incorrect labelling in our generated target data; this issue is confounded during evaluation as even though the bottom row shows that incorrectly labelled portion of “Construction” is predicted correctly by the network, this will count towards the incorrect labelling of the “Vehicle” class.

It is not surprising that larger objects such as “Construction”, “Vehicle” and “Vegetation” attain higher accuracies. As these objects are relatively large, the misalignment of the LiDAR scans with the RGB images leads to proportionately fewer incorrect labelled pixels. This is in contrast to smaller objects (such as a pole) which even a small misalignment can lead to the labelling of Pole-like where Construction exists and vice versa, for example. If a higher proportion of labels for smaller objects are incorrect, it is natural to expect inferior performance. A visual observation of the test results (such as in Figure 2) demonstrate that many of the labels which appear to be incorrectly labelled as the Construction class are in fact correctly predicted by the network.

A particular reason for the very poor performance of the Bike-like class is a nuance of the dataset used. The data is collected in the Oxford city centre which has a notorious large number of bicycles. Bicycles make up over 80% of the labels in the “Bike-like” class. Inspecting the RGB streams shows a very large number of these bikes fall into one of two categories. They are either alone and parked right up against objects belonging to the “Construction” class (i.e. walls, buildings and fences) or they are in groups on large bike racks. Furthermore due to the hollow nature of bicycles there are relatively few labels with bicycle labels near walls. In the first case, the result is that the network treats bicycle labels as noise, which are then ignored during training, and so are simply predicted to be the “Construction” class. In the second case the racks of bicycles appear similar to “Vegetation” in radar scans; for these reasons along with the more general ones outlined above, the performance in the “Bike-like” class is very poor.

Finally, in order to behave appropriately, mobile robots operating in urban environments are required to distinguish between a "Pole-like" object and a "Pedestrian". However from a single radar image, it is impossible, even for a human expert, to distinguish between them. It is expected then that within the "small" object classes, the network will confuse Pedestrians and Pole-Like objects.

VII. Conclusions

This work demonstrates a method for producing large amounts of training data for radar segmentation suitable for weak supervision. This provides a key benefit as it does not require expensive and laborious manual-labelling of radar scans. Furthermore, radar scans contain complex artefacts and so manual labelling would be limited to experts and might nevertheless be difficult. Unlike previous work on scene understanding using radar scans, no preprocessing of radar data is needed. The method of labelled data generation made use of publicly available models trained on publicly available datasets to provide the semantic segmentation of the environment observed in our own offline dataset, enabling methods for future work to utilize the depth of work in this related but not identical field. The relatively small network ($\sim$ 8M parameters) currently performs well with larger objects and demonstrates robustness to the imperfections of the data generation process. In the future we plan to retrain and test the system on the all-weather platform described in [27].

Acknowledgements

Prannay Kaul is supported by UK's Engineering and Physical Sciences Research Council (EPSRC) through the Centre for Doctoral Training (CDT) in Autonomous Intelligent Machines and Systems (AIMS) Programme Grant EP/L015897/1. Daniele De Martini is supported by the UK EPSRC programme grant EP/M019918/1. Matthew Gadd is supported by Innovate UK under CAV2 – Stream 1 CRD (DRIVEN). Paul Newman is supported by EPSRC Programme Grant EP/M019918/1. The authors would also like to acknowledge the support of the Assuring Autonomy International Programme, a partnership between Lloyds Register Foundation and the University of York.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] S. H. Cen and P. Newman, "Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[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)*. IEEE, 2019, pp. 298–304.

[4] R. Aldera, D. De Martini, M. Gadd, and P. Newman, "Fast Radar Motion Estimation with a Learnt Focus of Attention using Weak Supervision," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2019.

[5] 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 Conference (ITSC)*, Auckland, New Zealand, October 2019.

[6] D. Barnes, R. Weston, and I. Posner, "Masking by moving: Learning distraction-free radar odometry from pose information," in *Conference on Robot Learning (CoRL)*, 2019.

[7] Ș. Săftescu, M. Gadd, D. De Martini, D. Barnes, and P. Newman, "Kidnapped Radar: Topological Radar Localisation using Rotationally-Invariant Metric Learning," *arXiv preprint arXiv: 2001.09438*, 2020.

[8] T. Y. Tang, D. De Martini, D. Barnes, and P. Newman, "RSL-Net: Localising in Satellite Images From a Radar on the Ground," *arXiv preprint arXiv:2001.03233*, 2020.

[9] M. Adams, M. D. Adams, and E. Jose, *Robotic navigation and mapping with radar*. Artech House, 2012.

[10] S. Agarwal, J. O. D. Terrail, and F. Jurie, "Recent Advances in Object Detection in the Age of Deep Convolutional Neural Networks," *CoRR preprint arXiv: 1809.03193*, 2018.

[11] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably unknown: Deep inverse sensor modelling in radar," *arXiv preprint arXiv:1810.08151*, 2018.

[12] J. Lombacher, K. Laudt, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic radar grids," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1170–1175.

[13] N. Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, "Radar-based Feature Design and Multiclass Classification for Road User Recognition," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 779–786.

[14] Nicolas Scheiner, N. Appenrodt, J. Dickmann, and B. Sick, "Radar-based Road User Classification and Novelty Detection with Recurrent Neural Network Ensembles," *arXiv preprint arXiv:1905.11703*, 2019.

[15] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *2018 21st International Conference on Information Fusion (FUSION)*. IEEE, 2018, pp. 2179–2186.

[16] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[17] 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*. Springer, 2015, pp. 234–241.

[18] S. Hoermann, M. Bach, and K. Dietmayer, "Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2056–2063.

[19] L.-C. Chen, M. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for efficient multi-scale architectures for dense image prediction," in *Advances in Neural Information Processing Systems*, 2018, pp. 8699–8710.

[20] K. Shoemake, "Animating rotation with quaternion curves," in *ACM SIGGRAPH computer graphics*, vol. 19, no. 3. ACM, 1985, pp. 245–254.

[21] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[23] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[24] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset," *arXiv preprint arXiv: 1909.01300*, 2019.

[25] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1894–1900.

[26] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[27] S. Kyberd, J. Attias, P. Get, P. Murcutt, C. Prahacs, M. Towlson, S. Venn, A. Vasconcelos, M. Gadd, D. De Martini, and P. Newman, "The Hulk: Design and Development of a Weather-proof Vehicle for Long-term Autonomy in Outdoor Environments," in *International Conference on Field and Service Robotics (FSR)*, Tokyo, Japan, August 2019.