Keep off the Grass: Permissible Driving Routes from Radar with Weak Audio Supervision

David Williams*, Daniele De Martini*, Matthew Gadd*, Letizia Marchegiani†, and Paul Newman*
*Oxford Robotics Institute, Dept. Engineering Science, University of Oxford, UK.
{dw,daniele,mattgadd,pnewman}@robots.ox.ac.uk
†Automation and Control, Dept. Electronic Systems, Aalborg University, DK.
lm@es.aau.dk

***Abstract*— Reliable outdoor deployment of mobile robots requires the robust identification of permissible driving routes in a given environment. The performance of LiDAR and vision-based perception systems deteriorates significantly if certain environmental factors are present e.g. rain, fog, darkness. Perception systems based on Frequency-Modulated Continuous Wave scanning radar maintain full performance regardless of environmental conditions and with a longer range than alternative sensors. Learning to segment a radar scan based on driveability in a fully supervised manner is not feasible as labelling each radar scan on a bin-by-bin basis is both difficult and time-consuming to do by hand. We therefore weakly supervise the training of the radar-based classifier through an audio-based classifier that is able to predict the terrain type underneath the robot. By combining odometry, GPS and the terrain labels from the audio classifier, we are able to construct a terrain labelled trajectory of the robot in the environment which is then used to label the radar scans. Using a curriculum learning procedure, we then train a radar segmentation network to generalise beyond the initial labelling and to detect all permissible driving routes in the environment.**

***Index Terms*— radar, audio, terrain classification, weakly supervised learning**

I. Introduction

Safe navigation of intelligent mobile robots in unstructured and unknown outdoor environments (e.g. search and rescue, agriculture, and mining industry sectors) requires perception systems which deliver a detailed understanding of surroundings regardless of any environmental factor (e.g. weather, scene illumination, etc). In many environments, some terrains are unsuitable to traverse and so robust route identification is a key problem to be solved. To that end, a variety of sensor technologies have been used for solving related problems, including: cameras, LiDAR, sonar, audio, and radar.

LiDAR- and vision-based terrain classification systems are highly susceptible to inclement environmental or atmospheric conditions: heavy rain, fog, direct sunlight, and dust all greatly degrade the performance of these systems, thereby limiting their range of applications.

Frequency-Modulated Continuous Wave (FMCW) scanning radar, in contrast, operates robustly under such adverse conditions and additionally operates at ranges of up to many hundreds of metres – relaxing the maximum speed at which a robot can safely travel and facilitating longer planning horizons. Indeed, there is a burgeoning interest in exploiting FMCW radar to enable robust mobile autonomy, including ego-motion estimation [1]–[5], localisation [5]–[8], and scene understanding [9]–[11].

This project is supported by the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York, as well as UK EPSRC Programme Grant EP/M019918/1. Additionally, we would like to thank the Groundskeepers and Officers of the University Parks as well as our partners at Navtech Radar.

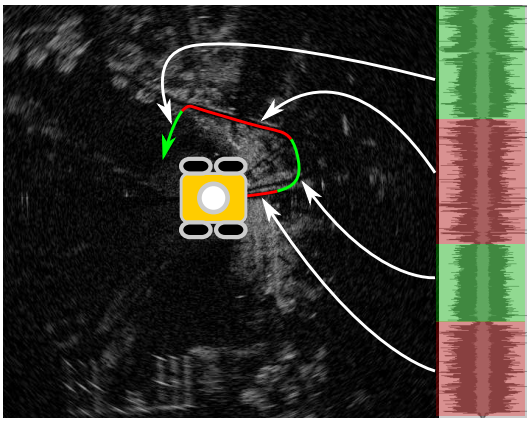


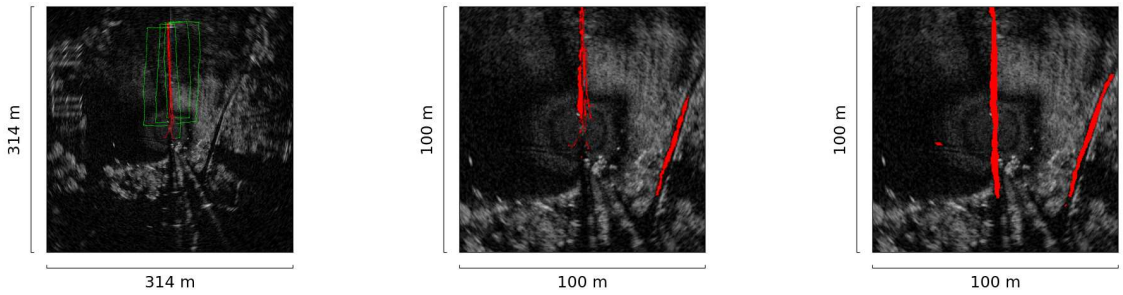
Fig. 1: Overview of the proposed system: audio is recorded and used to classify the terrain the robot is driving on – here gravel (red) and grass (green). Using odometry, the robot can paint this semantic information on top of the radar scan.

As a novel contribution to scene understanding with radar, this paper presents a system that detects permissible driving routes from raw radar scans. Specifically, it focusses on the methodology for the obtainment of labelling and a novel training procedure for the radar classifier.

Radar measurements are complex, containing significant multipath reflections, speckle noise, and other artefacts in addition to the radar's internal noise characteristics [12]. This makes the interaction of the electromagnetic wave in the environment more complex than that of time-of-flight (TOF) lasers. As obtaining a labelled radar dataset for supervision – with each scan annotated on a bin-by-bin basis – is challenging and time consuming, we propose an weakly-supervised framework using an alternative sensing modality: audio.

Audio-based terrain classifiers can be used to predict the permissibility of a driving route when the route is characterised by its terrain (e.g. grass, gravel, asphalt). Predicting terrain from audio is possible as each interaction between the robot and the ground has a terrain-specific audio signature.

Audio offers two advantages over other modalities,



(a) (b) (c)

Fig. 2: An example from the training dataset at three stages of the training process. (a) shows initial labelling (b) shows additional labels generated by stage one of the curriculum and (c) shows the final segmentation result.

e.g. vision-based systems: first, audio is invariant to scene appearance and less affected by weather conditions, providing more stable and predictable results; moreover, the use of microphones is advantageous as audio is a one-dimensional signal, easing the labelling process as the audio for each terrain can be collected separately.

Once the audio-based terrain classifier has been trained, we exploit it to weakly supervise the radar classifier training. Visual Odometry (VO) and GPS are used to trace the trajectory of the robot on the radar scan as if it were a canvas (see Figure 1) and each traversed bin is classified by the audio classifier.

II. Related Work

Mature techniques for identifying the driveable area of urban environments with cameras and LiDARs often learn to semantically segment the entire scene through the use of fully labelled datasets such as Cityscapes [13] or by weak supervision and demonstration as in [14]. In non-urban outdoor environments, path detection is closely related to the task of terrain classification [15]. For the environment in which our system was trained and tested, all permissible driving routes belong to one terrain class (gravel) and so for this application the tasks of permissible driving route identification and terrain classification are equivalent.

Vision-based terrain classification is perhaps the most traditional approach due to its associated intuitiveness and affordability. In [16], colour segmentation is employed to identify different terrains, while [15] performs both colour and texture segmentation for path detection. However in [16], problems arising due to variations in illumination are exposed. Although these problems are mitigable, when also paired with environmental factors such as fog, heavy rain and dust clouds, these systems alone seem unfit for robust autonomy.

LiDAR can be used to build successful terrain classifiers by observing the texture of the 3D point-cloud as seen in [17]. In low light conditions LiDAR works well, however it suffers greatly in the presence of rain and fog, limiting its applicability in much the same way as vision.

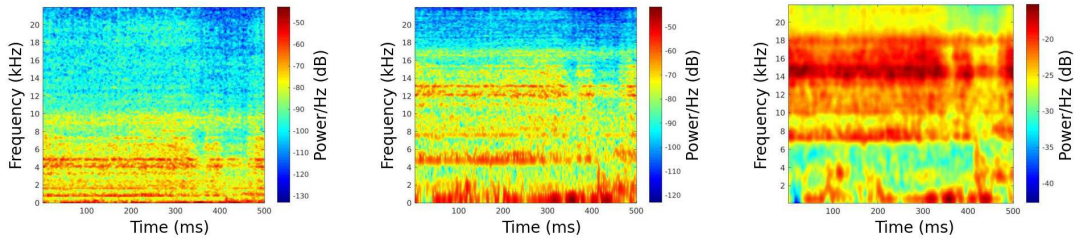
As mentioned in Section I, audio can also be used for terrain classification. Terrain-specific audio signatures are invariant to scene appearance and much less influenced by weather conditions compared with vision and LiDAR-based methods. The obvious disadvantage to this technique is that only the terrain the robot is currently operating on can be classified. As discussed in this paper, this characteristic can be leveraged for labelling purposes. [18] reports classification of nine different terrains with an accuracy of $99.41\,\%$ by leveraging advances in Deep Learning (DL) and using a Convolutional Neural Network (CNN) classifier. The audio features used for the CNN classifier were spectrograms generated with the Short Time Fourier Transform (STFT).

Despite the lower spatial resolution and compression of height information, in [19] it is shown in the context of a Simultaneous Localisation and Mapping (SLAM) system that while producing slightly less accurate maps than LiDARs, radars are capable of capturing details such as corners and small walls. This is reflected in literature as extensive research has been done using millimetre-wave radar systems for odometry, obstacle detection, mapping and outdoor reconstruction [1], [12], [20]. Radar is invariant to almost all environmental factors posed by even the most extreme environments, such as dusty underground mines, blizzards [21], [22]. Less work, however, has been carried out to investigate radar's performance on more comprehensive scene understanding tasks such as terrain classification or path identification. [23] presents an outdoor ground segmentation technique using a millimetre wave radar, however the chosen method limits its range of operation.

Perhaps most similar to our work is a visual terrain classifier which is also supervised by learned acoustic features presented in [24]. In our work, however, we focus on the usage of radar, which has advantages over vision in terms of robustness to both weather and illumination as well as sensor range. This work exposes at the same time challenges specific to the modality – especially the high sparsity of labelling. This is overcome with a stronger focus on the training procedure for the proposed network by explicitly promoting generalisation.

III. Methodology

Our method is based on our early investigation described in [9]. Learning to segment driveable routes in a radar scan – in a supervised manner – requires that routes in each scan are labelled. For a dataset of sufficient size (in the order of thousands of training examples), doing this by hand is a prohibitively time-consuming process. We therefore opt to weakly supervise the training of a radar-based segmentation network with an audio-based classifier that is



(a) Spectrogram (b) Mel-frequency spectrogram (c) Gammatonegram

Fig. 3: Visualisation of each time-frequency diagram used as feature representation for audio. Each diagram is generated from a clip length of $0.5\,\mathrm{s}$, shows frequencies up to half the sampling frequency and uses a bandwidth resolution of $100\,\mathrm{Hz}$.

trained independently of the radar-based classifier. Audio is collected for each terrain separately (making labelling trivial by tagging each sample as the whole sequence) and used to train the audio classifier for later use. Although the same labelling strategy could have been used to label radar directly, using audio as the labelling signal greatly increases the flexibility of the system by allowing multiple passes on different terrains in the same sequence and thus adding examples of both classes on single frames and at the same time lowering the sparsity of the labels.

Through the use of odometry and GPS, we obtain the data collection robot's timestamped trajectory in the environment. The audio terrain classifier is then used to accurately predict the terrain at each timestamp. By combining both, we produce a terrain-labelled trajectory of the robot in the environment (depicted in Figure 1) which is used as sparse labelling. For the purpose of segmenting paths in our environment, only the terrain labels denoting gravel are required.

A. Audio Classification

As audio is best interpreted as a sequence of frequencies correlated in time, we discuss its representation in the form of different types of spectrograms. As suggested in [18], spectrograms can be used as 1-channel images to feed into a CNN. This is effective as the success of CNN classifiers is in their ability to learn features automatically from data containing local spatial correlations. By assuming local spatial correlations in a spectrogram, the classifier recognises the temporal correlation of characteristic audio frequencies.

Our CNN classifier follows a standard architecture with several convolutional layers and max-pooling for downsampling.

For audio representation, we assess the performance of three types of spectrograms (results found in Section V). The representations considered are: Spectrograms, Mel-frequency spectrograms and Gammatonegrams (see Figure 3).

Spectrograms are the simplest time-frequency diagrams and are generated directly by the STFT. Mel-frequency spectrograms and gammatonegrams are motivated by the idea that the human auditory system does not perceive pitch in a linear manner. For humans, lower frequencies are perceptually much more important than higher frequencies and this can be represented in time-frequency representations. Gammatonegrams extend this biological inspiration, using filter banks modelled on the human cochlea and have been successfully used before in a robotics context [25].

The implementation used to generate both spectrograms and mel-frequency spectrograms is courtesy of VOICEBOX: Speech Processing Toolbox for MATLAB[1] and the MATLAB toolbox: Gammatone-like spectrograms[2] is used to generate gammatonegrams.

B. From Audio to Labelled Radar

In order to project terrain labels from audio into radar scans, we make use of the visual odometry estimate on the platform and GPS. VO produces a locally accurate, smooth trajectory and contains important orientation estimates. Although the estimates are locally accurate, they tend to drift over longer distances. In contrast, GPS measurements are globally accurate, but suffer from significant noise resulting in a non-smooth trajectory and contain low quality information about the orientation of the robot. In order to leverage the benefits of both techniques, we fuse these data streams using an Extended Kalman Filter (EKF).

Once the robot's trajectory has been generated, it is labelled using the audio classifier to predict the terrain for each timestamp. Finally, the labelled trajectory is fitted automatically to each radar scan using the position and orientation estimates from the EKF.

C. Radar Segmentation Training Procedure

The nature of the method used for collecting the labels means that the radar scans are both inexactly and sparsely labelled. The inexactness comes from measurement errors from the GPS and VO, and the sparsity comes from our inability to thoroughly traverse every driveable surface observed in the radar scans. This means that the training procedure must be designed such that the network can learn a more complex model than the labelling might immediately suggest.

To do this, data augmentation and a label propagation technique are used to design a two stage curriculum learning procedure. As described in [26], the idea of curriculum learning is that neural networks perform better when presented with the most understandable training examples first. This is done in the first stage by limiting the network's receptive field by only showing the network very small crops of the global scan. In this way, the network is restricted to simply learning

[1]Found at `www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. Produced by Mike Brookes, Dept. Electrical and Electronic Engineering, Imperial College in 1997.

[2]Found at `www.ee.columbia.edu/~dpwe/LabROSA/matlab/gammatonegram/`. Produced by Dan Ellis, Dept. of Electrical Engineering, Columbia University in 2009.

what a path looks like and is relieved of learning more complex concepts such as scene context. By comparison to the more difficult task of simultaneously segmenting multiple paths in the global scan, the network generalises much better on the simpler task of segmenting small crops (as suggested in [26]). For this reason, we are able to generalise beyond the initially incomplete labelling (see Figure 2). Before input to the network, crops are also flipped, rotated, elastically deformed and rescaled to expose the network to paths that are of different orientations, shapes and widths. This data augmentation promotes a broader understanding of how a path looks, thus assisting with generalisation.

Upon completion of the first stage, the network accurately segments small sections of paths contained in crops of the global scan (whether initially labelled or not) but is unsuited to segmenting the whole scan. The second stage of the curriculum is therefore to train the network to segment a whole scan containing multiple paths in one forward pass. By combining the predictions of the network from stage one and the original labelling, we obtain a more complete and exact set of labels from which the network can be trained to complete the more complex task. The idea of using a trained network's predictions to augment the labels is presented in a classification context in [27], however we adapt it to a segmentation context (described in Section V-B).

For the segmentation network, we chose a U-Net architecture [28], which has proven effective for segmentation of radar scans [2], [10]. A U-Net is a Fully Convolutional Network (FCN) containing downsampling and upsampling paths with skip connections between paths to propagate fine detail.

IV. Experimental Setup

This section discusses the platform and the dataset collected and used for training and testing of our system.

A. Platform and Sensors

A Clearpath Husky A200 robot was fitted with microphones and radar, for audio recording and route identification, and with cameras and GPS for odometry estimation. The audio data was recorded by using two Knowles omnidirectional boom microphones, mounted in proximity to the two front wheels, and an ALESIS IO4 audio interface, at a sampling frequency of $44.1\,\mathrm{kHz}$ and a resolution of 16 bits.

We used a Navtech CTS350-X FMCW scanning radar, mounted on top of the platform with an axis of rotation perpendicular to the road. The radar operates at a frequency of $76\,\mathrm{GHz}$ to $77\,\mathrm{GHz}$, yielding up to 3600 range readings, each constituting one of the 400 azimuth readings with a scan rotation rate of $4\,\mathrm{Hz}$. The radar's range resolution in short and long range configurations is $0.0438\,\mathrm{m}$ and $0.1752\,\mathrm{m}$ respectively, resulting in ranges of $157\,\mathrm{m}$ and $630\,\mathrm{m}$.

Images for VO were gathered by a Point Grey Bumblebee 2 camera, mounted facing the direction of motion on the front of the platform. GPS measurements were collected with a GlobalSat BU-353-S4 USB GPS Receiver.

B. Dataset

As discussed in Section III, audio was collected for each terrain separately. It was recorded from both microphones for $15\,\mathrm{min}$ per terrain class, corresponding to approximately 7200 spectrograms per class (using a clip length of $0.5\,\mathrm{s}$). Audio for grass and gravel terrains was collected in University Parks and the asphalt terrain in the Radcliffe Observatory Quarter.

Datasets for training and testing the classifier were collected with the radar in both the long range and short range configurations to ensure the network performs well regardless of specific radar configuration. We collected training data in two locations in University Parks, Oxford and testing data in two different locations in the same park. The audio classifier in combination with VO and GPS provides labelling for the training datasets. Figure 2 shows one location where the training dataset was collected comprises of two paths surrounded by grass. As the radar scan covers an area of $1\,587\,600\,\mathrm{m}^2$ in its longest range configuration, it is impractical to traverse every path observed by the radar. For this reason, we leave the side path untraversed (and therefore unlabelled), such that we can test the segmentation network's ability to generalise effectively.

V. Results

This section presents experimental evidence of the efficacy of our system.

A. Reliability of the Audio Supervisory Signal

An investigation was performed into the performance of the audio classifier using each different audio feature representation to determine which one would be used in the final classifier. In our experiments, the classifier is tested on a withheld testing dataset and predicts from three possible terrains: grass, gravel and asphalt. After averaging over multiple experiments, the accuracies for the spectrogram, mel-frequency spectrogram and gammatonegram were $98.5\,\%$, $98.8\,\%$, $99.4\,\%$ respectively (using a clip length of $0.5\,\mathrm{s}$). As the best performing feature representation, the gammatonegram was used to train the final audio terrain classifier.

Additionally, investigations into the audio clip length used to generate the gammatonegrams showed that the longer the clip length, the more accurate the terrain classifier. Whilst an intuitive result, this means a compromise between accuracy and system frequency is necessary. We chose a clip length of $0.5\,\mathrm{s}$ by balancing classification accuracy and other system frequencies (such as GPS update rate at $1\,\mathrm{Hz}$) to result in a classification frequency of $2\,\mathrm{Hz}$.

Lastly, the final audio terrain classifier was tested on a dataset where the robot dynamically traversed gravel and grass for $22\,\mathrm{min}$. Approximate hand-labels were generated by cross-referencing the predicted terrain at each of the 1320 GPS measurements with satellite imagery. Here, the audio terrain classifier performed the task with an accuracy of $98.4\,\%$.

B. Effective Supervision of Radar-only Segmentation

Firstly, a U-Net is trained on the training set shown in Figure 2(a) as stage one in the curriculum detailed in Section III.

Trained on the simple task of segmenting 64×64 crops out of a 512×512 scan, the network effectively learns not only to reproduce the labelling but also to segment paths unlabelled in our datasets (see Figure 2(b)).

To generate the labels for the previously unlabelled sections of scans, the radar scan is divided into a small sub-scans which are sequentially segmented by the trained network. To suppress spurious predictions, we randomly rotate each scan a small number of times and combine the predictions on each. Figure 2 shows an example of both the initial labelling and the generated labelling after stage 1.

Stage two of the curriculum involves fine-tuning the network with the newly generated dataset. We then test the network on datasets collected in two unseen locations with the radar in both long and short range configurations. Figure 4 shows both typical segmentations and some radar specific failure cases.

In both short and long range segmentations, the system is able to reliably detect driveable routes with a 360° field of view and up to hundreds of metres away. In Figures 4(d) and 4(g), we observe that paths approximately $100\,\mathrm{m}$ away and occluded by trees are accurately segmented in a way that would not be possible using any other sensor modality. Figure 4(a) shows the network segmenting around pedestrians and Figure 4(d) shows a consistent path detection behind occluding trees.

Figures 4(i), 4(n) and 4(p) show examples where occluded sections of the scan are misclassified as paths. This problem may be ameliorated by enforcing temporal consistency. In Figure 4(k), the vertical disjoint in the radar scan is misidentified as driveable path. This artefact arises due to the motion of the radar during scan formation, and can be fixed by motion correction. Finally, the network understandably doesn't predict through large occlusions, however could be achieved by fitting cubic curves between path segments as in [29].

The network correctly classified $98.8\,\%$ of pixels with an achieved IoU score of $39.8\,\%$ when evaluated on 25 hand-labelled unseen examples from the testing set. Compared with an IoU of $54.1\,\%$ achieved with cameras in [24] and considering radar's robustness to weather and illumination shows the feasibility of our method for all-weather scene understanding.

During inference, our U-Net runs at $330\,\mathrm{Hz}$ and uses less than $1\,\mathrm{GB}$ of GPU memory when processing 256×256 scans. We take this to be indicative that a CPU implementation may be feasible for closed-loop autonomy.

VI. Conclusions and Future Work

This paper presents a system that identifies permissible driving routes using scanning radar alone. With a specific focus on the methodology, the system is trained using an audio-leveraged automatic labelling procedure, followed by a curriculum designed to promote generalisation from sparse labelling. Qualitative results show that the network is capable of generalising effectively to the unseen testing set and to unlabelled areas of the training set. Quantitative results demonstrate the feasibility of our methodology for learning robust scene understanding from radar.

In the future, we plan to retrain and test the system on the all-weather platform described in [30], as part of closed-loop autonomy. Specifically, domains for deployment of this platform will be chosen to further explore the generalisibility of the presented system to other grassy environments as well as other driveable surfaces (e.g. asphalt). The proposed system will also be applied in off-road intelligent transportation contexts [31].

References

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

[2] 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*, 2019.

[3] 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*, 2019.

[4] 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.

[5] D. Barnes and I. Posner, "Under the radar: Learning to predict robust keypoints for odometry estimation and metric localisation in radar," *arXiv preprint arXiv: 2001.10789*, 2020.

[6] 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.

[7] M. Gadd, D. De Martini, and P. Newman, "Look Around You: Sequence-based Radar Place Recognition with Learned Rotational Invariance," in *IEEE/ION Position, Location and Navigation Symposium*, 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] D. Williams, D. De Martini, L. Marchegiani, and P. Newman, "Listening closely to see far away: Radar-based terrain classification from auditory signals," in *International Conference on Digital Image and Signal Processing (DISP)*, 2019.

[10] R. Weston, S. Cen, P. Newman, and I. Posner, "Probably Unknown: Deep Inverse Sensor Modelling Radar," in *2019 International Conference on Robotics and Automation*, 2019.

[11] P. Kaul, D. De Martini, M. Gadd, and P. Newman, "RSS-Net: Weakly-Supervised Multi-Class Semantic Segmentation with FMCW Radar," in *IEEE Intelligent Vehicles Symposium*, 2020.

[12] E. J. M. Adams, J. Mullane and B. Vo, *Robot Navigation and Mapping with Radar*. Artech House, 2012.

[13] 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 *IEEE conference on computer vision and pattern recognition*, 2016.

[14] D. Barnes, W. P. Maddern, and I. Posner, "Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy," *CoRR*, vol. abs/1610.01238, 2016.

[15] M. R. Blas, M. Agrawal, A. Sundaresan, and K. Konolige, "Fast color/texture segmentation for outdoor robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[16] P. Jansen, W. van der Mark, J. C. van den Heuvel, and F. C. Groen, *Colour based off-road environment and terrain type classification*. Piscataway, NJ: IEEE, 2005.

[17] M. Kragh, R. N. Jørgensen, and H. Pedersen, "Object detection and terrain classification in agricultural fields using 3d lidar data," in *International conference on computer vision systems*, 2015.

[18] A. Valada, L. Spinello, and W. Burgard, *Deep Feature Learning for Acoustics-Based Terrain Classification BT - Robotics Research: Volume 2*, A. Bicchi and W. Burgard, Eds. Springer, 2018.

[19] M. Mielle and a. L. A. J. Magnusson, Martin, "A comparative analysis of radar and lidar sensing for localization and mapping," in *European Conference on Mobile Robotics (ECMR)*, 2019.

[20] M. Heuer, A. Al-Hamadi, A. Rain, and M.-M. Meinecke, "Detection and tracking approach using an automotive radar to increase active pedestrian safety," in *IEEE Intelligent Vehicles Symposium*, 2014.

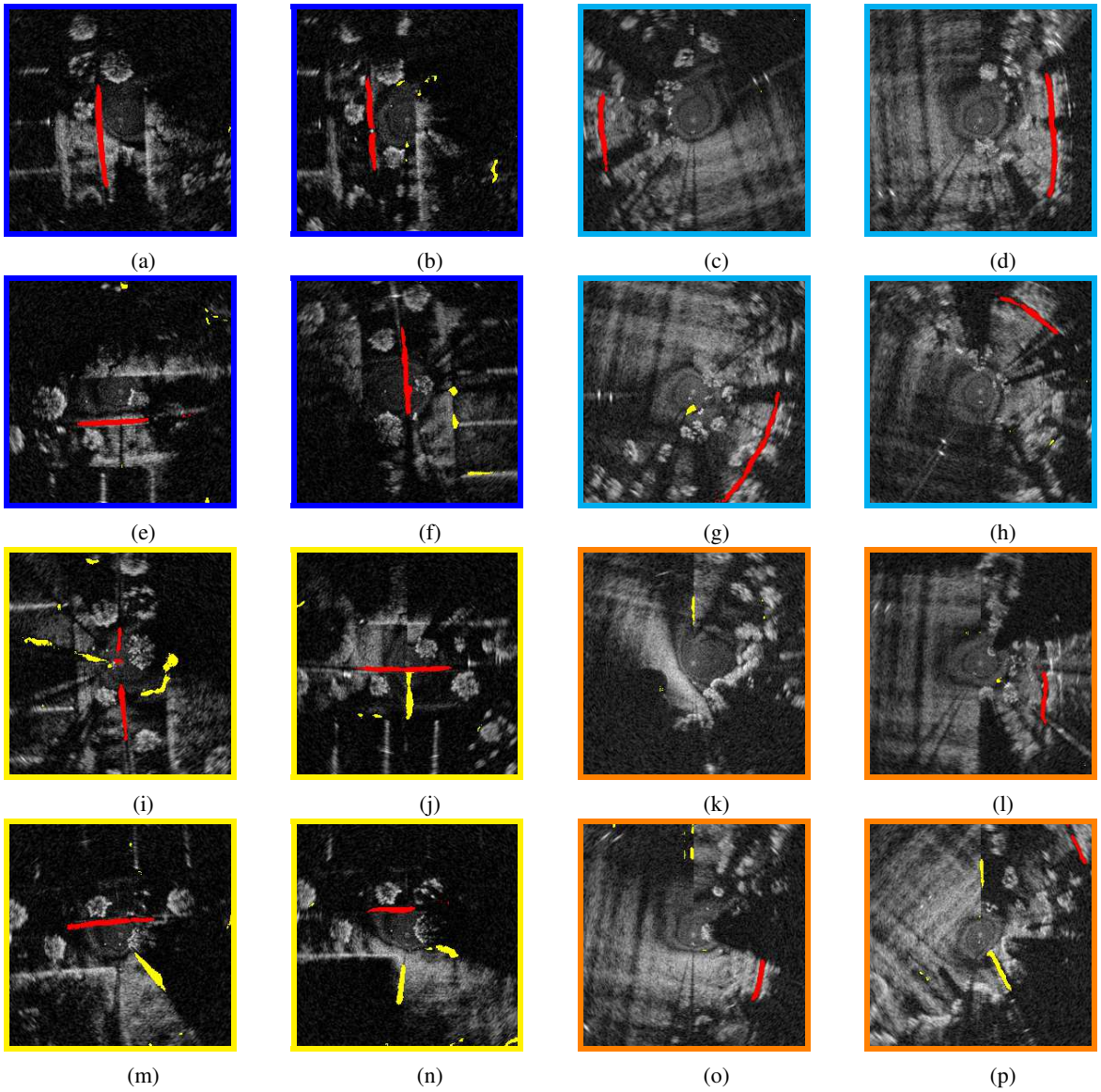


Fig. 4: U-Net segmentations of test examples in both short and long range configurations (with correct and incorrect predictions in red and yellow respectively). During inference, the dimensions of short and long range scans are $100\,\mathrm{m}$ and $400\,\mathrm{m}$ respectively (c.f. Section IV). Scans in the blue and yellow quadrants contain segmentations for the short-range configuration of the radar with yellow containing failure cases. Similarly, the cyan quadrant contains typical long-range examples, while the orange quadrant contains examples of long-range failure cases. See Section V-B for a discussion of these particular cases.

[21] G. Brooker, R. Hennessey, C. Lobsey, M. Bishop, and E. Widzyk-Capehart, "Seeing through dust and water vapor: Millimeter wave radar sensors for mining applications," *Journal of Field Robotics*, vol. 24, no. 7, 2007.

[22] A. Foessel, S. Chheda, and D. Apostolopoulos, *Short-range millimeter-wave radar perception in a polar environment*. Carnegie Mellon University, 1999.

[23] G. Reina, J. Underwood, G. Brooker, and H. Durrant-Whyte, "Radar-based perception for autonomous outdoor vehicles," *Journal of Field Robotics*, vol. 28, no. 6, 2011.

[24] J. Zürn, W. Burgard, and A. Valada, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *arXiv preprint arXiv:1912.03227*, 2019.

[25] L. Marchegiani and P. Newman, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," *ArXiv*, vol. abs/1810.04989, 2018.

[26] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," *Journal of the American Podiatry Association*, vol. 60, 2009.

[27] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imagenet classification through label progression," *CoRR*, vol. abs/1805.02641, 2018.

[28] 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*, 2015.

[29] T. Suleymanov, P. Amayo, and P. Newman, "Inferring road boundaries through and despite traffic," *2018 21st International Conference on Intelligent Transportation Systems*, 2018.

[30] 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*, 2019.

[31] M. Gadd, D. De Martini, L. Marchegiani, L. Kunze, and P. Newman, "Sense-Assess-eXplain (SAX): Building Trust in Autonomous Vehicles in Challenging Real-World Driving Scenarios," in *IEEE Intelligent Vehicles Symposium, Workshop on Ensuring and Validating Safety for Automated Vehicles (EVSAV)*, 2020.