High Resolution Radar Dataset for Semi-Supervised Learning of Dynamic Objects

Mohammadreza Mostajabi Ching Ming Wang Darsh Ranjan Gilbert Hsyu
Zendar Inc
{rmostajabi, jimmy, darsh, ghsyu}@zendar.io

Abstract

Current automotive radars output sparse point clouds with very low angular resolution. Such output lacks semantic information of the environment and has prevented radars from providing reliable redundancy when combined with cameras. This paper introduces the first true imaging-radar dataset for a diverse urban driving environments, with resolution matching that of lidar. To illustrate the need of having high resolution semantic information in modern radar applications, we show an unsupervised pretraining algorithm for deep neural networks to detect moving vehicles in radar data with limited ground-truth labels.

We envision that the details seen in this type of high-resolution radar image allow us to borrow from decades of computer vision research and develop radar applications that were not previously possible, such as mapping, localization and drivable area detection. This dataset is our first attempt to introduce such data to the vision community, and we will continue to provide datasets with improved features in the future.

1. Introduction

Camera, lidar and radar are the three main sensor types used in automotive for perception. Both lidar and radar are active range sensors. Between the two, radars are more robust under different lighting and adverse weather conditions. Radars are also much cheaper to manufacture. However, a typical automotive radar has 3-4 degree azimuth resolution and provides about ∼10,000 detected points per second. In comparison, a Velodyne VLP-16 lidar has 0.1 degree azimuth resolution and provides 300,000 points per second. This makes automotive radar quite inferior in both azimuth resolution as well as point density. The limited resolution of automotive radars stems from their limited antenna aperture. For 76-81GHz (automotive radar band) radars, a 2 meter long antenna aperture is required to achieve lidar like azimuth resolution. Under the current radar architecture, such an aperture is very difficult to achieve and the state-of-art radar has a 20cm aperture size.

For the dataset provided in this paper, we use an automotive radar system that is designed from the ground-up to achieve lidar-like resolution. By coherently combining information from multiple radars on the vehicle and utilizing accurate information about the vehicle's ego-motion, we can create a very long effective antenna aperture with a small physical aperture. This basic principle is known as synthetic aperture radar (SAR) and is commonly used in satellite and aerial radar imaging. By using this approach, we are able to image the static objects in the scene at 0.1 degree azimuth resolution and can resolve ∼1,000,000 points for a typical scene. Moreover, the SNR advantage of a large coherent aperture allows the radar to see very small objects and features that are not detectable using traditional radar. This advantage can be easily seen from Figure 1. A recorded video can be seen at [3]. The dataset will be available for download at zendar.io/dataset.

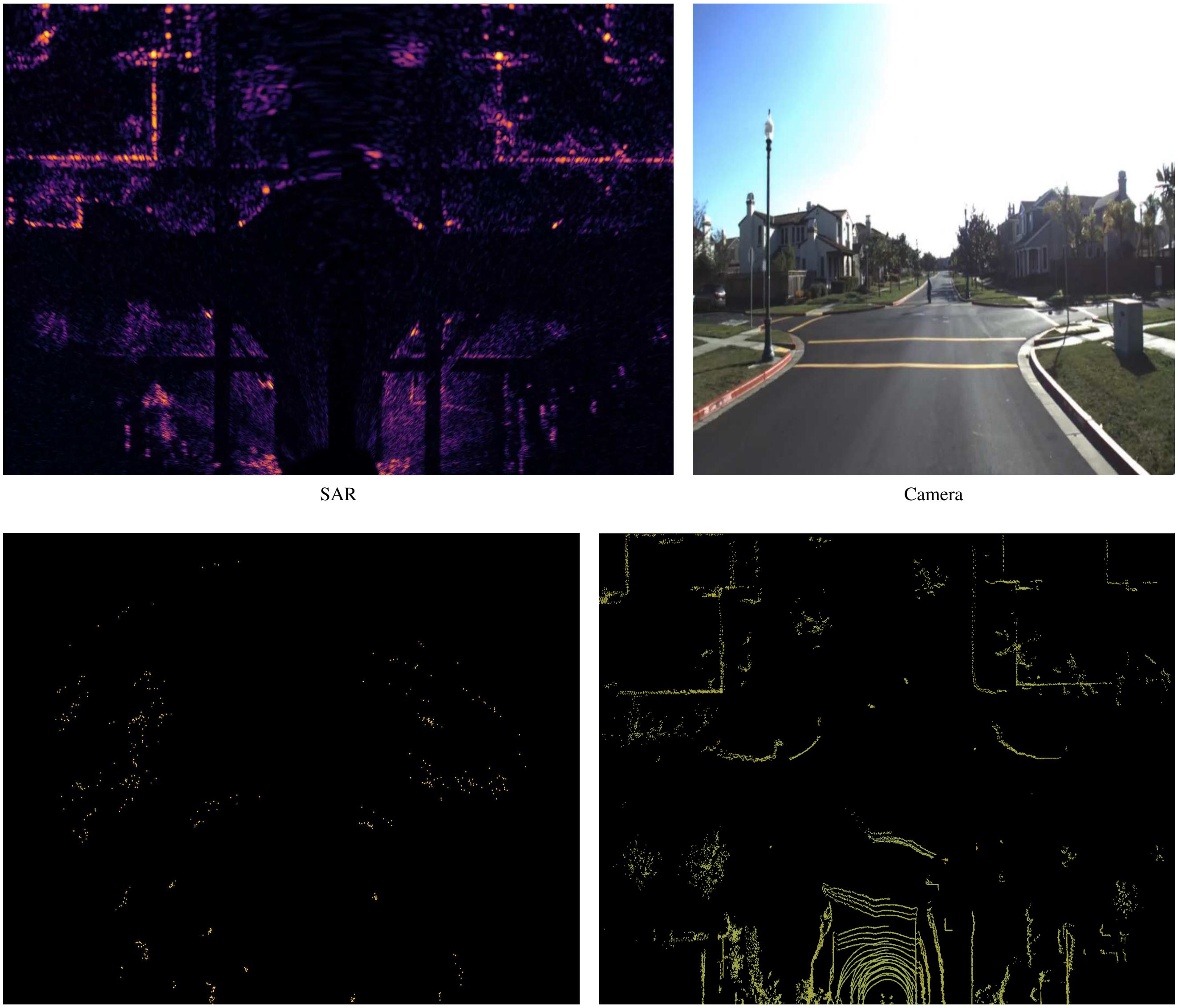
1.1. Synthetic Aperture Radar

Synthetic aperture radar (SAR) is a set of related techniques for coherently combining radar returns from a moving radar over some segment of the the radar's path to create a high-resolution two-dimensional image of the scene. The specific set of radar positions used to create the image is called the *synthetic aperture*. The concept of SAR is usually credited to the mathematician Carl Wiley of Goodyear Aerospace in 1951 and was originally devised as a method for imaging the Earth's surface from radar sensors moving at high altitude, first on aircraft and soon after on spacecraft [5].

While a detailed discussion of SAR imaging can be found in [5], we briefly describe some of the most salient characteristics of SAR images here.

1.1.1 Intrinsic Resolution

It is convenient for discussion to take the two axes of the image to be "range," the distance from the radar, and "az-



SAR

Camera

Traditional Radar Point Cloud

lidar

Figure 1: The same scene seen from camera image, bird’s eye view of SAR, traditional radar point cloud and lidar.

imuth,” the horizontal angular coordinate running perpendicular to the range. With these definitions, the intrinsic resolution of the SAR system can be decomposed into range and azimuth components. The intrinsic range resolution of the system is directly proportional to the carrier frequency bandwidth of the transmitted chirp signal and may be expressed in units of distance as

$$\Delta R = \frac{c}{2B},$$

where c is the speed of light and B is the carrier frequency bandwidth, which is at most 4 gigahertz for our data.

The intrinsic azimuth resolution of the system is directly proportional to both the absolute carrier frequency and the length of the synthetic aperture and may be expressed in radians as

$$\Delta\theta = \frac{c}{2fL\sin\theta},$$

where f is the central carrier frequency, L is the synthetic aperture length, and θ is the azimuth coordinate, with $\theta = \pm\pi/2$ being opposite directions parallel to the synthetic aperture. This means that in theory, azimuth resolution can be pushed arbitrarily high by using very long synthetic apertures. In practice, the effective synthetic aperture length is limited by (1) the extents of the physical radar beam, meaning that a particular area in the scene may not be visible from all positions along the radar's path, and (2) errors in ego-position estimation, which hamper the ability to coherently combine radar returns over very long synthetic apertures.

1.1.2 Stationary and Moving Objects

The SAR approach requires accurate relative distance measurement between the radar and the object. If all features in the scene are stationary, then only vehicle ego-motion needs to be measured. This is achieved by using the GPS/IMU system available on the vehicle. In order to image objects that are moving fast (i.e. comparable to ego speed of vehicle), the object's motion needs to be estimated as well. One possible way is to use a traditional radar signal processing pipeline based on short-aperture range-Doppler-DOA (direction of arrival) processing in parallel with the SAR processing pipeline.

When a moving object's path is not compensated prior to SAR imaging, two things occur:

1. Its image *defocuses*, i.e., loses resolution and SNR.
2. Its image may be displaced in azimuth.

Both effects are proportional to the ratio of the object's uncompensated speed to the ego-vehicle's speed. Thus, since only stationary objects image cleanly with high resolution, accuracy, and SNR, to a rough approximation, a SAR image without object motion compensation may be considered as a "static-scene" datum, i.e., a representation of the stationary elements of the scene only. Because our SAR data stream does not incorporate object motion compensation, this heuristic can be exploited to help distinguish static from moving objects, as described in 2.3.

1.1.3 Automotive SAR

SAR has been used for decades, first by the military and later civilian operators, for aerial surveillance and remote sensing. More recently, SAR has been extended to the automotive setting (e.g., [6]), but the benefits provided by the high-definition radar imaging enabled by SAR in this setting are only beginning to be explored. Thus, while there already exists a body of high-altitude SAR data (e.g., [1]), this is not the case for street-level SAR. One of our key contributions is thus the publication of a high-quality street-level SAR dataset consisting of complex urban imagery.

1.2. Contributions

In this paper, we describe a semi-supervised method for the detection of moving objects in the range-Doppler-DOA domain. We have accurately labeled 11,000 moving cars with their positions on range-Doppler maps and SAR images to train and evaluate the performance of radar-based moving object detection algorithms.

Our contributions:

- First high resolution radar image dataset of complex urban driving scenes.
- Time synchronized radar data cube, radar images with camera images and lidar overlay.
- Novel semi-supervised training of deep neural networks for radar dynamic object detection.
- Over 11,000 moving cars labeled in 27 diverse scenes with over 40,0000 automatically generated labels of moving cars to train and evaluate performance of moving object detection algorithms on complex and diverse urban driving scene scenarios.

2. Semi-supervised Dynamic Object Detection

2.1. Range-Doppler Preprocessing

A traditional radar processing pipeline assumes the world is made of point-like reflective objects. The signal reflected from such an object has a very simple mathematical model parametrized by:

1. *range*, the distance from the radar;
2. *Doppler velocity*, the rate of change of range; and
3. *direction of arrival (DOA)*, the direction vector from the radar to the object, which is often further decomposed into *azimuth* (horizontal) and *elevation* angles.

The signal reflected from the world is then modeled as the linear superposition of the reflections from all the individual objects [8].

The data as received from the hardware is a time series of receiver antenna excitation levels, indexed by *slow time* (which counts individual chirps), *fast time* (which counts ADC samples within a chirp), and the *virtual antenna index*, which enumerates the physical array elements on the sensor (more precisely, active pairs of transmitting and receiving antennas). However, in practice, it is convenient to take a discrete time chunk of the data and *compress* it along the two time dimensions, transforming those axes into the Doppler and range axes, respectively.

Because this "compression" essentially amounts to a two-dimensional discrete Fourier transform and is thus reversible [8], this range-Doppler-domain format may also be described as a "raw" radar format.

Range-Doppler compression performs two important functions:

1. It decouples the compressed range and Doppler axes from the uncompressed DOA axis. This enables, for example, nonparametric detection over range and Doppler combined with parametric estimation of DOA. For this to be effective, the intrinsic range and Doppler resolutions of the system need to be sufficient, since nonparametric detection will only report detections to within one range-Doppler resolution cell. (Superresolution is possible, but this again requires joint estimation of all parameters for all objects contributing to a given resolution cell.)

2. The response of an individual object of interest is generally highly localized in range and Doppler, unlike in the uncompressed signal model. If the range and Doppler axes have resolution sufficient to distinguish typical objects of interest like cars and pedestrians, as they do in our case, then the localized range-Doppler structure of these objects can be exploited by sophisticated detection and classification algorithms. This is an active topic of current and future research, and we will describe this subsequently. This localized structure in the range-Doppler domain can be seen in Figure 2.

2.2. CNN Detector

One candidate algorithm that can model the multidimensional structure of moving objects in radar data is a convolutional neural net (CNN). Previous work using CNN on radar data uses simple and shallow network structure [15]. The networks have limited receptive field size and fail to extract features in different scales and distortions. Our goal is to use a network with large receptive field size, supporting multiple scales and utilizing temporal information. The major challenge with training such a network is the scarcity of labeled data. Although there are a few open source radar datasets [4], [13], the labels are on the processed sparse point clouds. This is the first labeled dataset on raw radar data and SAR point cloud.

2.3. Unsupervised Dynamic Object Detection

In order to generate enough labels for pretraining, we combine the high-resolution static-scene images with the output of a traditional radar pipeline. Specifically, the traditional radar pipeline outputs a sparse list of points, each of which belongs to either a moving object or a stationary object. By projecting the radar points into the static scene (i.e., SAR image), we can distinguish a moving object from a stationary one. Here, we exploit the heuristic that the SAR image cleanly images primarily the static elements of the scene and thus argue that the points detected on moving objects are likely to be those that project to low-SNR regions of the SAR image. However, note that this heuristic is very approximate, since moving objects do pollute the SAR images. Thus, the output of this step should be considered as fairly noisy in subsequent processing.

All the potential moving object points are projected back into the raw radar data and segmentation is performed around each point. Segmented clusters are subsequently labeled.

The pretraining is then supplemented with supervised fine-tuning. The results are discussed in Section 4.

3. Dataset

Our dataset consists of the following data streams. The elements of each data stream are timestamped by a common GPS clock, so the timestamps may be used to synchronize and multiplex all the streams. Note that the elements of the radar and lidar streams all aggregate data over a time interval. By convention, these stream elements are timestamped by the GPS time at the center of this time interval.

3.1. Raw Radar Data Cube

This is the radar data format commonly used in automotive radar processing. The full time series is segmented into fixed-length chunks and range-Doppler-compressed as described in section 2. Thus, each chunk may be described as a three-dimensional complex-valued array with the axes

1. range sample, with 512 elements,

2. Doppler sample, with 256 elements,

3. antenna array element (transmitter-receiver pair), with 4 elements (one transmitter and four receivers) arrayed uniformly along azimuth only. (Thus, the system has no elevation resolution.)

This is illustrated in Figure 4.

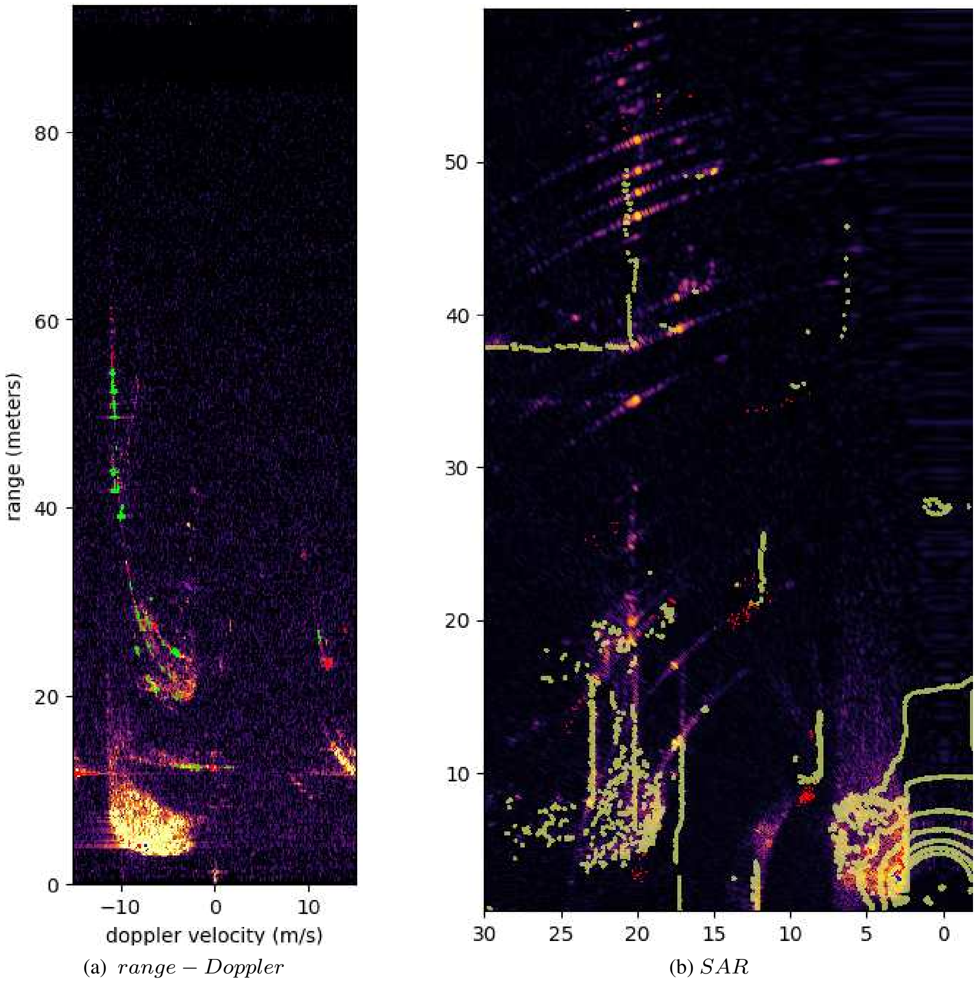
Along the range dimension, the radar maximum range is set to 90m and the range resolution is 18cm.

Each radar cube is timestamped by the center time of the pre-Fourier-transform time series.

3.2. High Resolution Static Scene (SAR)

Each image is created using SAR backprojection [7] from a two-meter segment of radar data (i.e., two-meter *synthetic apertures*). Note that the time-extent of the data used to create this image is thus inversely proportional to the speed of the vehicle at the time of collection, in contrast to the raw radar cubes, which have fixed extents in time. The width of an image pixel is 4 centimeters along both axes.

This stream carries a fixed frame rate of 10 frames per second. Thus, the two-meter synthetic apertures used to image each frame will overlap when the collection vehicle travels slower than 20 meters per second.



(a) *range − Doppler*

(b) *SAR*

Figure 2: Automatic dynamic object extraction: (a) Green points on the range-Doppler map are stationary detections masked out by SAR. Red detection are dynamic detections. Blue detections are uncertain, meaning that the cluster that they belong to contains both stationary and dynamic detections. (b) Bird's eye view of SAR point cloud with lidar overlay (yellow points) and dynamic detections from range-Doppler (red points).

Each image is tagged with a timestamp that corresponds to the middle of the synthetic radar aperture.

3.3. Tracklog

The tracklog consists of the vehicle pose stream recorded by the VectorNav VN200, the onboard GPS/IMU. Each entry in the tracklog represents the rotation and translation from the IMU sensor frame (forward, right, down) to the Earth-centered, Earth-fixed (ECEF) global frame. The extrinsic pose of each sensor is also provided to convert into the sensor-centric frame.

3.4. Camera Image

The reference camera image is taken by ImageSource (model number DFK-33GX265e) color industrial camera. The frame rate is set to 24Hz.

3.5. Lidar Point Cloud

This dataset uses a Velodyne VLP-16 lidar to provide ground truth for the dynamic objects. For every image in the static-scene radar image stream, the corresponding lidar points within the raw radar data time range are projected into the radar image frame. The lidar point cloud frame rate is 10Hz, identical to the high-resolution static-scene radar stream.

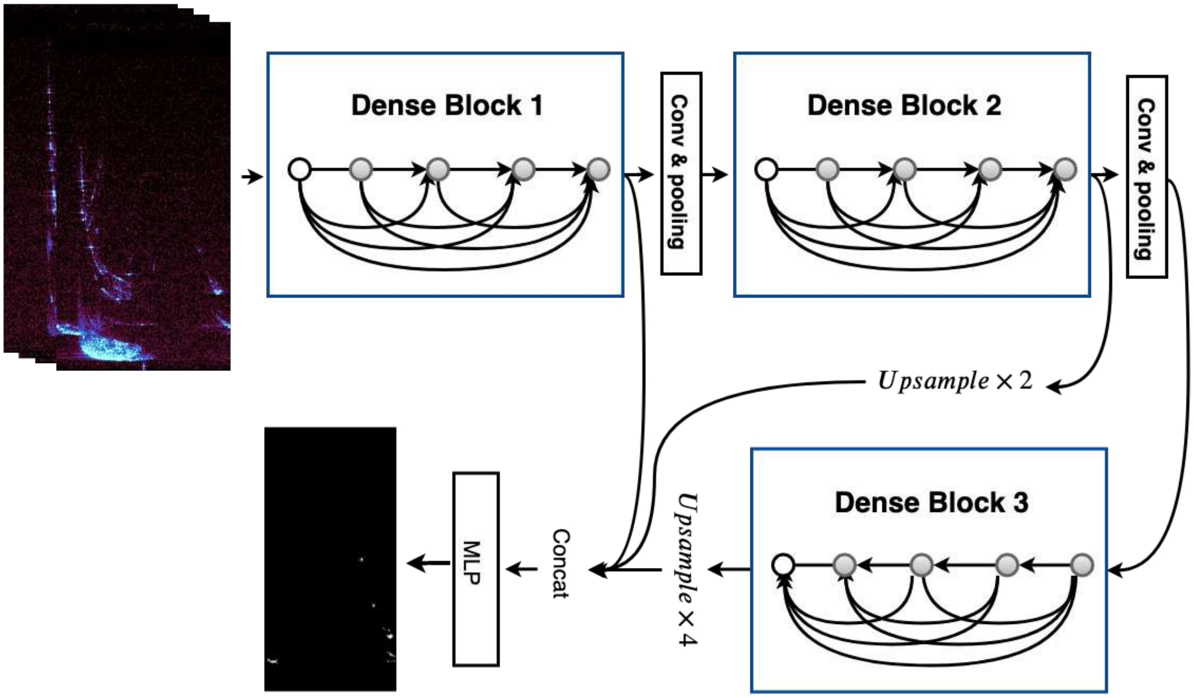


Figure 3: Illusteration of Zoom-out DenseNet with three dense blocks. Input to the network is a sequence of range-doppler maps over time. Network output is dynamic objects mask.

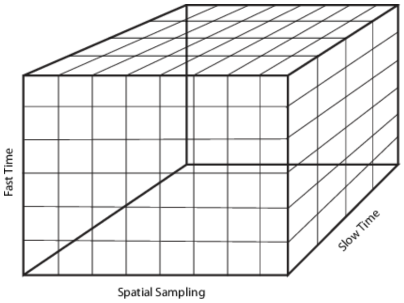


Figure 4: Radar data cube

4. Experiments

The performance of our baseline moving-object detector models is evaluated on the test set of our dataset. The test set has over 1300 labeled moving cars in complex urban driving scenarios. The training set contains 10,000 labeled and 400,000 automatically generated labels of moving cars. CNN based experiments are done in PyTorch [2], using the Adam [11] update rule when training networks.

Unsupervised labels are generated based on the unsupervised dynamic object detection procedure 2 with Constant False-Alarm Rate (CFAR) detector. Models trained on unsupervised labels use a batch size of 8 and learning rate of 10^{-4} for 100 epochs. These models are then fine-tuned on the labeled data by end-to-end training with learning rate of 10^{-5} for 80 epochs.

4.1. Evaluation Metrics

In order for a detection to be considered true positive, a common measure is to compute the area of the overlap between the predicted bounding box and the ground truth bounding box. The predicted bounding box is true positive if the area of the overlap is greater than a threshold. In our dataset, we have used lidar to annotate bounding boxes on the SAR image. Since we don't expect point-wise radar detections and lidar to have exact returns from the same parts of a moving car, we consider a predicted bounding box to be true positive if the center of the bounding box is closer than half the length of an average car, i.e. 2.5 meters, to the center of the ground truth bounding box. We use average precision (AP) measure to evaluate the dynamic object detection performance. Range-based AP is computed in terms of the area under the precision-recall curve at different ranges.

4.2. Dynamic Object Extraction

A sequence of range-Doppler frames provides a rich temporal feature for dynamic object detection as dynamic objects have a different movement pattern compared to stationary objects over time in the range-Doppler map. There-

Method	unsupervised	supervised	AP near	AP mid	AP far
CFAR	-	-	65	43	13.3
FPNs ResNet-34	-	✓	66.3	56	22.8
FPNs ResNet-34	✓	✓	73.1	63.4	23.9
zoom-out DenseNet-37	-	✓	67	59.8	28.3
zoom-out DenseNet-37	✓	✓	77.2	68.1	29.2
zoom-out DenseNet-67	-	✓	74.8	58.8	29.6
zoom-out DenseNet-67	✓	✓	76.6	67.6	29.9

Table 1: Moving car detection results on the test set.

fore, we want to utilize a CNN architecture that has a large receptive field, extracts features at multiple scales and utilizes temporal information.

We choose Feature Pyramid Networks (FPNs) and zoom-out CNN architectures [12, 14] as a primary basis for our baseline experiments due to the ability of these architectures to extract feature at multiple spatial scales. The input to the dynamic object detection CNN is a couple of range-Doppler maps concatenated across the time axis. The training is done end-to-end to extract dynamic objects. Extracted dynamic objects are then clustered and projected into the SAR image.

We have trained a FPNs model with a 34-layer ResNet [9, 12] backbone and two zoom-out models with DenseNet [10] backbone with growth rate of 32 on training subset of our dataset. The first zoom-out model consists of a 37-layer DenseNet with two dense blocks. A dense block consists of multiple convolutional layers where each layer is connected to every other layer. All layers within the same dense block operate on the feature maps of the same spatial size. The second zoom-out model has a 67-layer DenseNet backbone with four dense blocks. Figure 3 illustrates zoom-out DenseNet architecture with three dense blocks.

4.3. Analysis

Table 1 presents the results of dynamic car detection with CFAR and CNN detectors. We have reported the range based area under the curve results. Near range is from 0 to 15 meters, mid and far ranges are 15-30 and 30-60 meters respectively. Regardless of the detector type, we apply our proposed unsupervised dynamic object detection procedure to the output of the detector. We will miss some of the dynamic objects if they are not detected by the CFAR or CNN detector. False positive cases can rise from failure in creating a perfect SAR mask or detecting noisy sidelobes in the range-Doppler map. The other source of false positives is imperfect estimation of direction of arrival (DOA), especially for far objects. This will lead to improper positioning of detected stationary points on the SAR image and potentially masking them using the static SAR mask. As it is shown, CNN models significantly outperform the CFAR-based detector. The signal-to-noise ratio is much lower for far objects, which makes them difficult to detect. Even under low signal-to-noise ratio for far objects, CNN models produce significantly higher AP comparing to the CFAR model at test time. Unsupervised pretraining of the CNN models consistently improves AP across different models and different ranges.

5. Conclusion

We introduce a high-resolution imaging radar dataset collected from challenging urban driving environments. With our proposed dataset, the computer vision community now has access to raw radar data and high-resolution SAR images along with reference lidar and camera data. In this paper, we demonstrate the benefit of using SAR images to automatically generate labels to pretrain a deep CNN and show improved performance compared to traditional radar detection methods.

It is our belief that this type of high-resolution image data enables many radar applications that are previously not possible, such as scene segmentation and object detection. We plan to release sample data of this type in the future.

References

[1] Polsarpro data sources. https://earth.esa.int/web/polsarpro/airborne-data-sources.

[2] PyTorch. https://github.com/pytorch/pytorch.

[3] *Dual Corner Radar Video*, 2019. `https://www.youtube.com/watch?v=ERduNkYokjU&feature=youtu.be`.

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[5] Ian G. Cumming and Frank H. Wong. *Digital Processing of Synthetic Aperture Radar Data: Algorithms and Implementation*. Artech House Remote Sensing Library. Artech House, 2005.

[6] Juergen Dickmann, Nils Appenrodt, Jens Klappstein, Hans-Ludwig Bloecher, Marc Muntzinger, Alfons Sailer, Markus Hahn, and Carsten Brenk. Making bertha see even more: Radar contribution. *Access, IEEE*, 3:1233–1247, 01 2015.

[7] Armin Doerry, Edward Bishop, and John Miller. Basics of backprojection algorithm for processing synthetic aperture radar images. 02 2016.

[8] F. Engels, P. Heidenreich, A. M. Zoubir, F. K. Jondral, and M. Wintermantel. Advances in automotive radar: A framework on computationally efficient high-resolution frequency estimation. *IEEE Signal Processing Magazine*, 34(2):36–46, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *CVPR*, 2017.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[12] Tsung-Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *CVPR*, 2017.

[13] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. *European Radar Conference*, 2019.

[14] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. *CVPR*, 2015.

[15] K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer, and B. Yang. Deep learning-based object classification on automotive radar spectra. *IEEE Radar Conference (RadarConf)*, 2019.