

LiRaNet: End-to-End Trajectory Prediction using Spatio-Temporal Radar Fusion

Meet Shah^{*1}, Zhiling Huang^{*1}, Ankit Laddha^{*1}, Matthew Langford¹, Blake Barber¹, Sidney Zhang¹, Carlos Vallespi-Gonzalez¹, Raquel Urtasun^{1,2}

¹Uber Advanced Technologies Group, ² University of Toronto
`{meet.shah, zhiling, aladdha, mlangford, bbarber, sidney, cvallespi, urtasun}@uber.com`

Abstract: In this paper, we present LiRaNet, a novel end-to-end trajectory prediction method which utilizes radar sensor information along with widely used lidar and high definition (HD) maps. Automotive radar provides rich, complementary information, allowing for longer range vehicle detection as well as instantaneous radial velocity measurements. However, there are factors that make the fusion of lidar and radar information challenging, such as the relatively low angular resolution of radar measurements, their sparsity and the lack of exact time synchronization with lidar. To overcome these challenges, we propose an efficient spatio-temporal radar feature extraction scheme which achieves state-of-the-art performance on multiple large-scale datasets. Further, by incorporating radar information, we show a 52% reduction in prediction error for objects with high acceleration and a 16% reduction in prediction error for objects at longer range.

Keywords: Radar, Spatio-Temporal Sensor Fusion, Trajectory Prediction

1 Introduction

Trajectory prediction plays a pivotal role in the success of self-driving vehicles (SDVs) in dynamic and interactive environments. The SDV can plan a safe path by taking into account the future position of objects to preemptively avoid potentially harmful interactions. It involves understanding each actor’s motion and intention as well as the complex interactions with other actors and the environment. Due to its importance for SDVs, a plethora of work in trajectory prediction has been undertaken [1, 2, 3, 4, 5, 6].

Recently, end-to-end approaches that directly predict trajectories from raw sensor data have gained traction [7, 2, 8, 4, 5, 9, 10] due to their efficiency and high performance. These approaches typically exploit lidar as well as HD maps for this task. Lidar captures positional information about objects, while HD maps provide scene context and a prior on possible paths. To accurately predict future motion, these methods need to implicitly understand the dynamics of objects in the scene. They do so by relying on a sequence of position observations from lidar over time, increasing the reaction time of the vehicle. This can be problematic in certain critical situations where accurate, low-latency motion prediction is needed to avoid an imminent collision.

Radar, on the other hand, provides instantaneous velocity and positional information. Radar has been deployed in numerous applications in the past half century ranging from subterranean to space¹ and more recently, has been deployed as a primary sensing modality in commercial ADAS systems². Given the ubiquity of radar sensing in many domains for trajectory prediction, it is surprising that it has not featured more prominently in end-to-end prediction systems for self-driving. However, the benefit of radar in terms of instantaneous velocity measurement also comes with challenges. Typical automotive radars have lower resolution and provide point clouds with higher position uncertainty (Fig. 1) than typical lidars. Furthermore, it only provides the radial component of the object velocity as depicted in Figure 1. As a result, most existing approaches [11, 12, 13] utilize radar by heavily

^{*}Equal Contribution

¹<https://en.wikipedia.org/wiki/radar>

²https://en.wikipedia.org/wiki/Advanced_driver-assistance_systems

relying on late fusion. They leverage the object’s position for associating radar points to the object, and the radial velocity measurement from these associated radar points to estimate object velocity. Therefore, they require a multi-stage (detect-track-predict) method to utilize a temporal sequence of radar and for predicting trajectories.

In contrast, we directly do an early fusion of a temporal sequence of radar data along with lidar and map for trajectory prediction. We propose a novel end-to-end approach, called LiRaNet, to fuse dynamic information from radar with lidar and HD maps. In particular, we present an *early fusion* method that effectively combines radar data with other sensors at feature level. To overcome the sparsity, positional uncertainty and partial velocity measurements of radar, our method fuses a sequence of past radar data to learn spatio-temporal features on a bird’s eye view (BEV) grid using a graph. These dynamics-rich radar based features are combined with lidar and map features containing detailed geometric information to accurately predict the future positions of objects.

Our proposed method provides state-of-art trajectory prediction performance on two large-scale datasets. We demonstrate that by effectively exploiting radar information our approach provides substantial improvements on three different trajectory prediction formulations. Furthermore, we showcase that adding radar significantly improves trajectory prediction in safety and latency critical scenarios where lidar-only methods struggle due to sparse points and rapidly changing dynamics.

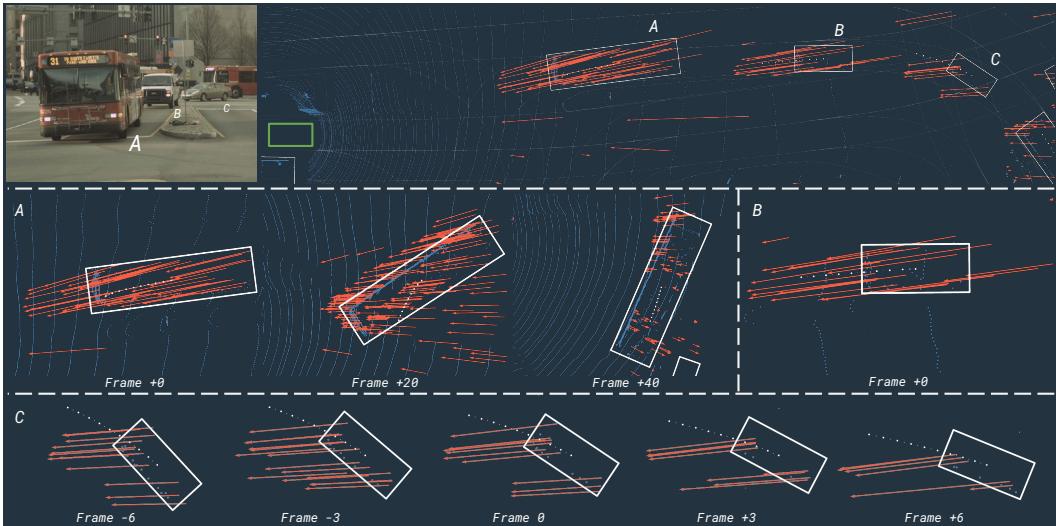


Figure 1: Challenges and Benefits of Radar: An example scene from X17k in bird’s eye view where lidar points (light blue) and radar point velocities (orange) are visualized with labels (white) for current, past and future frames. Vehicle A is a turning bus that has multiple radar points across frames. By effectively combining them over space and time a full 2D velocity and turning rate can be recovered. Vehicle B shows the high positional noise that inherently comes with radar. Vehicle C shows a case with sparse lidar points where implicitly associating them across time can be challenging. However, radar points present around C can add context for the model to detect and predict the trajectory. More examples (with model predictions) where radar data can positively complement lidar can be found in Fig 3.

2 Related Work

Trajectory Prediction: Classical methods [1, 14, 3, 15, 16, 17, 6] utilize a multi-stage detect-track-predict paradigm for trajectory prediction. In the first stage, sensor data is used to detect objects [18, 19, 20, 21, 22, 23, 24] at each time step independently. These objects are then tracked across time to generate a temporal sequence of object positions. The sequence of positions are then used for trajectory prediction. By dividing the problem into stages, independent progress can be made for each step and different sensors can be fused at different stages. However, such approaches suffer from cascading errors and high latency since multiple models are run sequentially. This motivates the recent usage of end-to-end single stage methods that can potentially improve performance.

The seminal work of [7] proposed a single stage end-to-end approach for predicting trajectories directly from raw lidar data using a BEV grid. This approach is further improved in subsequent

works by incorporating HD map [2], reasoning about interactions [4, 8], using a better network architecture [5, 8] and range view based lidar representation [9, 10]. As compared to multi-stage approaches, these methods are faster, easier to maintain in a production environment and have higher performance [4]. They can easily take advantage of sensor information and context of other objects for predictions. In this paper, we propose to add radar along with lidar and HD maps for improving end-to-end trajectory prediction.

Multi-Hypothesis Trajectory Prediction: Many single stage methods produce a single deterministic trajectory [2, 8] or a single distribution such as Gaussian [4] or Laplacian [9, 10, 5] for each waypoint of the trajectory. However, to ensure safe and efficient operations, an autonomous vehicle is required to anticipate a multitude of possible behaviors of actors in the scene. This has motivated a plethora of work [25, 6, 26, 17, 27, 15] in probabilistic multi-hypothesis trajectory prediction. We incorporate the mixture distribution formulation of [25] and [6] in our proposed method and demonstrate improvements by incorporating radar.

Radar for Self Driving: Automotive radars have several desirable properties such as cost effectiveness and resilience to extreme weather. These have led to it being a widely used sensor for ADAS and has recently motivated diverse research efforts to utilize radar information for self driving. However, most existing methods utilize radar only for improving perception [28, 13, 12, 29, 30, 31, 11] which has an indirect impact on trajectory prediction through a multi-stage pipeline. In comparison, our proposed approach directly improves trajectory prediction. Methods utilizing radar data can be divided into two major groups: tracking based and learning based.

Classical tracking based approaches [28, 13, 12] fuse radar data with other sensors using filtering techniques such as Kalman Filters to create tracks. Dynamics information in each track can be integrated forward for trajectory prediction. However, data driven approaches based on deep learning can easily outperform [26, 6] such simple prediction methods. Recently, several learning based approaches using radar data for self driving have been proposed. UNets are used in [32, 33] to learn semantic segmentation of BEV grid using only radar data. Radar and Camera are fused for 3D object detection in [29, 30, 34]. Nonetheless, rich velocity information provided by radar is ignored in all of these approaches. More recently, joint 3D detection and velocity estimation is performed in [31] using raw Range-Azimuth-Doppler tensor and in [11] by fusing lidar and radar. In contrast to these, we propose to utilize radar for trajectory prediction in an end-to-end learning based method.

3 Multi-Sensor Trajectory Prediction

We propose a novel approach for end-to-end trajectory prediction by fusing velocity-rich information from radar with other sensor data. We use a Bird’s Eye View (BEV) grid which is commonly used to fuse multiple sensors [35] as well as temporal sequences of 3D sensors [7]. Thus, we propose to learn spatio-temporal features for BEV grid cells using a sequence of past radar data. Figure 2 shows an overview of our proposed approach. Each input sensor modality is first processed individually to generate multiscale BEV feature volumes. Feature volumes from multiple modalities are then merged together using concatenation and then passed through a backbone network to generate features for each cell in the grid. These are then fed into our object detection and trajectory prediction headers to obtain predictions.

In the following sections, we first describe data from radar sensors and the challenges associated with it. Next, we describe our proposed approach to learn features for BEV grid cells using radar. Then, we describe other sensor representations and network architecture used to fuse all the modalities. Finally, we discuss the end-to-end training scheme of our proposed network.

3.1 Radar Sensor Data

Radar observes the environment by transmitting and receiving radio waves. The raw data cube generated by the radar is generally too large to be practically relayed off-sensor or processed without specialized hardware. Off the shelf automotive radar sensors overcome this by detecting peaks in the raw radar data cube using a Constant False Alarm Rate (CFAR) [36] algorithm, and only reporting these peaks (the rest of the data is assumed to be noise). Given this processing, we consider these peaks, called radar points, for feature learning.

Until now, radar data has not been used for end-to-end trajectory prediction in self driving due to a few key difficulties and shortcomings. Current-generation automotive radar sensors can: contain 2-3

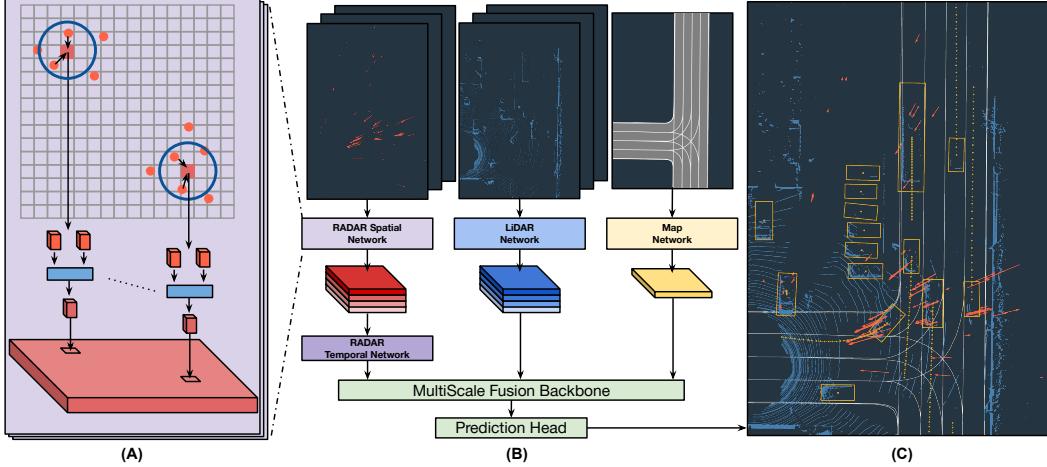


Figure 2: LiRaNet overview: Our radar feature extraction network **(A)** extracts spatio-temporal features from raw radar points in 2 steps: (1) for each frame we create a graph between the BEV grid cells and radar points to learn spatial features of each cell using a non-rigid convolution, (2) these spatial features are further fused temporally by stacking across channel dimension and using an MLP to get a radar feature volume. This feature volume is then fused with feature volumes from other sensors and fed to a joint perception-prediction network **(B)** which produces detections and their future trajectories. An example prediction for a scene from X17k can be seen in **(C)**.

orders of magnitude fewer points relative to lidar, have angular accuracies 1-2 orders of magnitude worse relative to lidar, and while they do provide a direct observation of velocity, they can only observe velocity in the radial direction. All of these challenges complicate the fusion of lidar and radar. However, intuitively the complementary radial velocity measurement of radar should be able to help in improving the estimate of object dynamics. In the following section we propose a method to overcome these challenges to effectively learn features for BEV grid cells and improve prediction.

3.2 Spatio-Temporal Feature Learning Using Radar

Multiple radar sensors are usually placed on a vehicle to achieve a 360° view. We create a radar sweep such that it contains the measurements from all radars in a time interval. We assume we are given a sequence of M such radar sweeps as input. All radar points are transformed into the coordinate frame associated to the most recent sweep. For each radar point i in the sweep m , we calculate a feature vector f_i^m containing its 2D position (x_i^m, y_i^m), radar cross-section (RCS) (e_i^m) and ego-motion compensated radial velocity (expressed as a 2D vector in the vehicle coordinate frame) ($v_{x,i}^m, v_{y,i}^m$). We propose to use two separate modules to learn features from radar points for BEV cells: spatial and temporal. The spatial module works on a each sweep individually and provides robustness to sparsity and position errors observed in radar points. The temporal learning module provides robustness to radial velocity observation by using a sequence of sweeps to implicitly recover full 2D velocity of objects.

Spatial Feature Learning: The goal of this module is to extract BEV features for each cell in the BEV grid using a single sweep of radar points. The grid is chosen to be centered at the ego-vehicle position at the most recent sweep. In comparison to lidar, radar is very sparse and has high positional noise. Thus, directly discretizing the radar x,y coordinates at the same resolution as lidar for feature extraction does not work well (see Sec 4). Therefore, we propose the following novel feature learning method.

We first select an appropriate resolution for feature learning. This resolution is usually lower than the resolution for lidar due to sparsity in radar points. We then use a learnable operator to extract the features at this resolution. Unfortunately, we cannot use the standard convolutional operators since the source domain (unordered points) is not a grid like structure. Therefore, we use the recently proposed parametric continuous convolutions [37] for computing the BEV cell features. Parametric convolution generalizes the standard convolution operator to non-grid like structure and can handle different input and output domains with pre-defined correspondence between them. In particular we use the more expressive adaptation of [35].

In our case, the input domain consists of radar points and output domain consists of BEV cells. Therefore, for each cell j we can calculate the features \mathbf{h}_j^m for the sweep m as (see Figure 2):

$$\mathbf{h}_j^m = \sum_{i \in \mathbf{A}_j^m} g^m(\mathbf{f}_i^m \oplus (\mathbf{x}_i^m - \mathbf{x}_j^m)) \quad (1)$$

where \mathbf{A}_j^m is the set of associated radar points, \mathbf{x}_i^m is the 2D coordinates of the associated radar point, \mathbf{x}_j^m is the 2D coordinate of the BEV cell's center, \oplus denotes the concatenation operation, \mathbf{f}_i^m is the feature vector for the radar point and $g^m(\cdot)$ is an multi-layer perceptron (MLP) with learnable weights shared across all the cells. We calculate \mathbf{A}_j^m , using nearest neighbor algorithm with a distance threshold. By using a threshold larger than the size of a cell, our method compensates for positional errors in radar.

Temporal Feature Learning: The goal of this module is to combine the spatial features for all the sweeps. For each cell j , we calculate the final spatio-temporal feature vector (\mathbf{h}_j) by concatenating the per sweep features $\mathbf{h}_{j,m}$ and using an MLP to combine them.

3.3 Lidar and HD Map Feature Learning

We assume that we are given L past lidar sweeps and an HD map of the surroundings. We represent the HD map [2, 5] as a rasterized BEV image containing geometric lane information. We employ a lightweight network to learn map features for BEV cells using the rasterized image. We represent lidar as an BEV occupancy grid [7]. Since lidar is dense and has very low position noise, we can assume that the lidar point coordinates (x, y, z) are the true position of the object from which these points are reflected. Thus we can directly discretize the coordinates to calculate occupancy features for each BEV cell. For points in each sweep, the x and y dimensions are discretized to create a BEV grid and the z dimension is discretized to create a multi channel binary occupancy feature for each cell in the grid. These temporal, per sweep occupancy features are concatenated and used as input to a lightweight network to learn spatio-temporal lidar features for the BEV cells. Additional details about feature learning from lidar and HD map data can be found in the supplementary material.

3.4 Backbone Network Architecture

We use the same backbone network architecture as [38, 5]. In particular, we use the multi-scale inception module [38] as the main building block for extracting and combining multiple scale features. The inception block contains three branches, each with a down-sampling ratio of 1x, 2x and 4x. We use 3 sequential inception blocks and then use a feature pyramid network to compute the features for each BEV cell. We concatenate features from all modalities at multiple scales and provide it as input to the first inception block. The multiscale features for each modality are computed using a single strided convolution. A dense single stage convolutional header [18] is used for detecting objects from the pyramid features. For trajectory prediction, we extract a large rotated ROI [4, 5] of $40mx40m$ centered at the object to learn actor centric features which are then used to predict the final trajectory [5, 25, 6].

3.5 End-to-End Training

We jointly train the proposed method using a multi-task loss defined as a weighted sum of the detection and trajectory loss: $\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda \mathcal{L}_{traj}$.

Detection Loss (\mathcal{L}_{det}) is a multi-task loss defined as a weighted sum of classification and regression loss: $\mathcal{L}_{det} = \mathcal{L}_{det}^{cls} + \alpha \mathcal{L}_{det}^{reg}$. We use focal loss [39] to train classification of each BEV cell for being at the center of an object class. We parameterize each object i by its center (x_i, y_i) , orientation (θ_i) and size (w_i, h_i) . We use smooth L1 loss to train regression parameters of each object.

Trajectory Loss (\mathcal{L}_{traj}): To evaluate the generality of our proposed method, we consider multiple different trajectory formulations: single-hypothesis [5] and multi-hypothesis [25, 6]. We consider each waypoint at time t of a trajectory j to be a 2D Laplace distribution parameterized by its position (x_j^t, y_j^t) and scale $(\sigma_{j,x}^t, \sigma_{j,y}^t)$. We use the sum of KL divergence [40] between the ground truth and predicted distribution for all the waypoints as our regression loss for trajectory. In case of single hypothesis prediction, the prediction loss only contains the regression component, $\mathcal{L}_{traj} = \mathcal{L}_{traj}^{reg}$. However, in multiple-hypothesis prediction, it contains both regression and classification, $\mathcal{L}_{traj} = \mathcal{L}_{traj}^{reg} + \beta \mathcal{L}_{traj}^{cls}$. We use cross-entropy loss for learning the confidence of each predicted hypothesis.

Method	Radar	60% Recall				80% Recall			
		All		Moving		All		Moving	
		ADE ↓	FDE ↓						
SpAGNN [4]	✗	-	145	-	-	-	-	-	-
LaserFlow [9]	✗	-	153	-	-	-	-	-	-
RV-FuseNet [10]	✗	-	123	-	-	-	-	-	-
MultiXNet* [5]	✗	62	113	141	287	68	122	157	314
LiRaNet (Ours)	✓	56	102	127	263	63	106	136	278

Table 1: **Trajectory prediction performance on nuScenes[41].** Both ADE and FDE are in cm.

Method	Radar	All		Moving	
		ADE ↓	FDE ↓	ADE ↓	FDE ↓
MultiXNet* [5]	✗	35	56	154	326
LiRaNet (Ours)	✓	33	52	135	285

Table 2: **Trajectory prediction performance on X17k.** Both ADE and FDE are in cm.

4 Experimental Results

Datasets: We evaluate our approach on two large-scale autonomous driving datasets: nuScenes and X17k. nuScense is publicly available and X17k is collected in-house in dense urban settings. For each lidar sweep in these datasets, we find the nearest (in time) radar sweep to synchronize them. X17k contains an order of magnitude more data than nuScenes with ~ 2.5 million sweeps for training and $\sim 600k$ for validation.

For nuScenes, we use the official ROI of 100m square centered at the SDV. Whereas, for X17k dataset we use an ROI of square of 100m in front of the vehicle. For training and testing the trajectory prediction we use the sensor data from past 0.5 seconds and predict the trajectory for 3 seconds in future. We use objects of vehicle class for evaluation in both datasets. For additional training details, refer to the supplementary material.

Metrics: For evaluating trajectory predictions for each object we use: Average Displacement Error (ADE) and Final Displacement Error (FDE) defined as following:

$$\text{ADE} = \frac{1}{4} \sum_{t=0}^3 \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2 \quad \text{FDE} = \|\mathbf{x}_3 - \hat{\mathbf{x}}_3\|_2 \quad (2)$$

where $\hat{\mathbf{x}}_t$ is the predicted position at time t (in s) and \mathbf{x}_t is the ground truth position. To calculate these metrics, we consider a detection to be true positive if it has an IoU ≥ 0.5 with the ground truth. We use the same recall value of 80% for comparing methods with different detection sets. We train and test on the vehicle class in both datasets.

4.1 Comparison with State of the Art

Single Hypothesis Methods: Table 1 shows the comparison of our approach with previous single-hypothesis approaches on nuScenes. Note that here we additionally report prediction results on 60% recall points since all existing method only report results in this setting and for fair comparison we compare our method in their setting. We use a state-of-the-art model MultiXNet [5] as our lidar and map based baseline and show that adding radar significantly improves results on nuScenes 1.

In particular we get a substantial $\sim 13\%$ reduction in the prediction error at 3s for moving objects. This demonstrates that the proposed approach of learning feature is effectively able to exploit radar velocity information. Furthermore, using radar we achieve state-of-art results on this dataset.

Table 2 demonstrates the impact of incorporating radar into MultiXNet [5] on our in-house large-scale dataset (X17k). Here we see a similar $\sim 13\%$ reduction in prediction error for moving objects. This dataset is collected in dense urban settings which inherently contain a lot of static vehicles ($\sim 85\%$ of all the labels). Overall improvement on all vehicles is therefore lower since the presence of a relatively high proportion of static vehicles masks the improvement on the moving vehicles.

Multiple Hypothesis Methods: We incorporate anchorless MTP [25] and anchor-based MultiPath [6] multi-hypothesis prediction formulations in our proposed network and study the impact of adding radar. We use 16 hypotheses for both MTP and MultiPath. Table 3 shows the results on moving objects. We can see that adding radar data helps on both formulations and in both datasets. We

Method	Radar	ADE ₁ ↓	ADE ₃ ↓	ADE ₅ ↓	ADE ₁₀ ↓	ADE ₁₅ ↓
MTP [25]	✗	162 184	142 137	122 116	94 86	76 77
MTP(Ours)	✓	150 181	112 132	95 112	70 86	65 77
MultiPath [6]	✗	206 180	122 108	102 95	96 90	76 89
MultiPath(Ours)	✓	168 163	95 101	81 91	75 87	74 85

Table 3: **Trajectory prediction performance on moving objects using multi-hypothesis formulations.** ADE_k (in cm) denotes the minimum of the ADE over top k hypothesis on nuScenes (black) and X17k(blue)

K	d	ADE ↓	FDE ↓	Res.	ADE ↓	FDE ↓	History	ADE ↓	FDE ↓
1	1	140 142	295 310	0.125x	136 143	280 311	0s	157 154	314 326
1	10	136 135	278 285	0.25x	140 142	295 310	0.1s	138 142	285 307
1	25	137 138	279 302	0.5x	141 146	291 316	0.2s	136 141	281 310
1	+Inf	140 140	284 301	1x	146 150	297 324	0.5s	136 135	278 285
2	+Inf	142 145	290 315						

(a) Impact of Graph Parameters (b) Impact of Grid Resolution (c) Impact of Radar History

Table 4: **Ablation studies:** We study the impact of different parameters of our feature learning scheme for moving objects on nuScenes and X17k for the single hypothesis setting. K and d denote the number of neighbors and the distance threshold (in number of cells) respectively. Res denotes the resolution as ratio of the size of the radar feature extraction grid to the lidar occupancy grid. History denotes the size of temporal context for radar. Using this study, we chose 0.25x resolution, $K = 1$, $d = 10$ and a radar history of 0.5s as our final model.

also see that the performance improvement is higher when we chose a lower number of hypotheses. This clearly shows that adding radar is not only able to produce better hypotheses but is able to do so with higher confidence as can be seen in Fig. 3.

4.2 Comparison on Different Object Characteristics

Radar provides improvements in cases where lidar has difficulty in learning object dynamics. Figure 4 shows the improvements in FDE of objects over different object characteristics on X17k in the single hypothesis model. In the following, we analyze some of the factors where trajectory prediction is particularly challenging.

Position and Scene based Factors: We first look at objects based on the number of lidar points on them and their distance from the SDV. In Figure 4(b), we see significant gains in prediction performance as the number of lidar points on an object decreases. Implicit association of sparse point clouds across time for learning the dynamics is a difficult problem. Therefore, having a direct measurement of velocity reduces the error by $\sim 16\%$ in the case of objects with as few as 1-10 lidar points. Similarly, since distance from SDV is inversely related to the number of lidar points, Figure 4(a) shows that the improvements from radar increase at longer ranges.

Object Dynamics based Factors: We now look at objects based on their speed and acceleration. Figure 4(c) shows that radar can improve predictions on objects regardless of speed. The very low speed ($\leq 4m/s$) objects are usually easier to associate across time, so the positional information of lidar alone is usually enough for learning dynamics. Figure 4(d) shows the improvements of radar based on acceleration of the object. We can clearly see that radar improves predictions by a large margin as the acceleration increases. Since it takes more observations of lidar (positional information) to learn acceleration as compared to radar (velocity information), it reduces the error by more than 50% for objects with very high acceleration of greater than $5m/s^2$.

4.3 Ablation of Radar Feature Learning

We have proposed multiple design strategies to address the challenges for fusion of radar with other sensors. In the following section, we analyze the importance of those design choices.

Graph Parameters: Table 4a shows the impact of various parameters of the graph for association of BEV cells with radar points. From the table we can see that using a reasonably wide area to find the neighbors can improve the results significantly by $\sim 8\%$ as compared to sparse association ($K = 1$ and $d = 1$). These results clearly show that for achieving good performance, it is necessary to extract features for BEV cells with points which do not directly fall in them. Additionally, we observe that having higher number of neighbors does not provide additional value due to the sparsity.

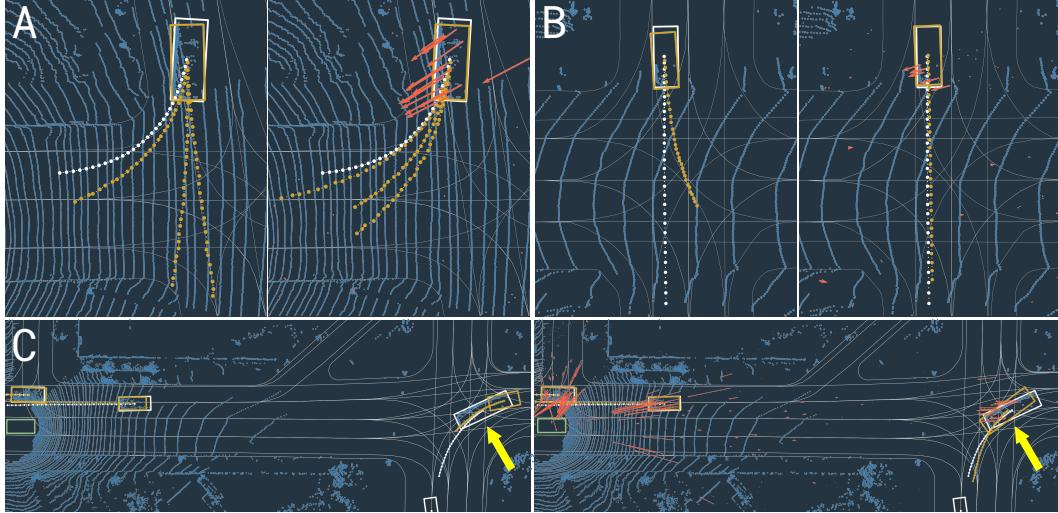


Figure 3: **Qualitative Improvements:** These three example scenes from X17k show improvements with the addition of radar data. Labels are shown in white and predicted output is shown in gold. In (A), the vehicle just started slowing down (has high deceleration) for the turn. This change in velocity can be hard to measure with historical lidar alone so the lidar model still gives a high probability for the mode which goes straight. With radar however, all of the modes correctly indicate that the vehicle is turning. (B) shows a case where lidar is sparse due to occlusion which leads to an incorrect estimation of the object turning. Since radar shows approximately zero radial velocity, with radar the model can correctly infer that it is going straight at high velocity. In (C), the vehicle pointed at by the yellow arrow is too far away the lidar model to correctly estimate the shape, heading or velocity. However, with radar radial velocity observations all can be correctly estimated.

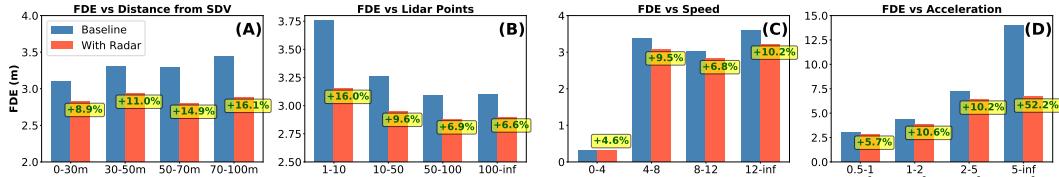


Figure 4: **Results by different object characteristics:** We show relative improvement (%) on X17k in FDE metric (m) by adding radar. (A) is binned by distance to AV, (B) is binned by number of lidar points, (C) is binned by speed, (D) is binned by acceleration (the absolute value). We can infer that adding radar significantly improves over baselines in several challenging scenarios.

BEV Resolution: Table 4b shows the impact of BEV cell resolution for learning radar based features. Due to sparsity of radar we can see that selecting a resolution that is 0.25 times lower than lidar can significantly improve the impact of adding radar on both datasets.

Radar History: Table 4c shows the impact of temporal radar context. Since radar only provides radial component of velocity, multiple observations on an object are needed to implicitly learn the tangential component. For objects with a high number of radar points, it may be possible to learn this from a single sweep. On objects with sparse radar points, however, these observations have to come from past frames. This clearly shows value from increasing temporal context for radar. The performance on X17k improves more than nuScenes due to the longer average range of objects in X17k where the sparsity of radar is more of an issue.

5 Conclusion

In this work we proposed using instantaneous velocity information of radar for improving the end-to-end trajectory prediction. To this end, we proposed an efficient spatio-temporal radar feature extraction scheme which achieves state-of-the-art performance on multiple large-scale datasets. We further demonstrated the advantages leveraging rich dynamic information from radar in challenging scenarios where lidar struggles. Our approach can increase overall robustness of end-to-end trajectory prediction systems, particularly in scenarios with rapidly changing dynamics where lidar-only baselines struggle and potentially reduce harmful interactions.

References

- [1] N. Djuric, V. Radosavljevic, H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, and J. Schneider. Motion prediction of traffic actors for autonomous driving using deep convolutional networks. *arXiv preprint arXiv:1808.05819*, 2018.
- [2] S. Casas, W. Luo, and R. Urtasun. IntentNet: Learning to predict intention from raw sensor data. In *Proc. of the CoRL*, 2018.
- [3] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proc. of the IEEE CVPR*, 2019.
- [4] S. Casas, C. Gulino, R. Liao, and R. Urtasun. Spatially-aware graph neural networks for relational behavior forecasting from sensor data. *arXiv preprint arXiv:1910.08233*, 2019.
- [5] N. Djuric, H. Cui, Z. Su, S. Wu, H. Wang, F.-C. Chou, L. S. Martin, S. Feng, R. Hu, Y. Xu, A. Dayan, S. Zhang, B. C. Becker, G. P. Meyer, C. V-Gonzalez, and C. K. Wellington. Multi-xnet: Multiclass multistage multimodal motion prediction, 2020.
- [6] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov. MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*, 2019.
- [7] W. Luo, B. Yang, and R. Urtasun. Fast and furious: Real time end-to-end 3D detection, tracking and motion forecasting with a single convolutional net. In *Proc. of the IEEE CVPR*, 2018.
- [8] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li. Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *Proc of the IEEE CVPR*, 2020.
- [9] G. P. Meyer, J. Charland, S. Pandey, A. Laddha, C. V-Gonzalez, and C. K. Wellington. Laserflow: Efficient and probabilistic object detection and motion forecasting. *arXiv preprint arXiv:2003.05982*, 2020.
- [10] A. Laddha, S. Gautam, G. P. Meyer, and C. Vallespi-Gonzalez. Rv-fusenet: Range view based fusion of time-series lidar data for joint 3d object detection and motion forecasting. *arXiv preprint arXiv:2005.10863*, 2020.
- [11] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *Proc. of the ECCV*, 2020.
- [12] H. Cho, Y.-W. Seo, B. V. Kumar, and R. R. Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *2014 IEEE ICRA*, 2014.
- [13] D. Göhring, M. Wang, M. Schnürmacher, and T. Ganjineh. Radar/lidar sensor fusion for car-following on highways. In *The 5th International Conference on Automation, Robotics and Applications*, pages 407–412. IEEE, 2011.
- [14] J. Hong, B. Sapp, and J. Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *Proc. of the IEEE CVPR*, 2019.
- [15] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker. DESIRE: Distant future prediction in dynamic scenes with interacting agents. In *Proc. of the IEEE CVPR*, 2017.
- [16] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proc. of the IEEE ICCV*, 2019.
- [17] C. Tang and R. R. Salakhutdinov. Multiple futures prediction. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, 2019.
- [18] B. Yang, W. Luo, and R. Urtasun. PIXOR: Real-time 3D object detection from point clouds. In *Proc. of the IEEE CVPR*, 2018.
- [19] G. P. Meyer, A. Laddha, E. Kee, C. V-Gonzalez, and C. K. Wellington. LaserNet: An efficient probabilistic 3D object detector for autonomous driving. In *Proc. of the IEEE CVPR*, 2019.
- [20] G. P. Meyer, J. Charland, D. Hegde, A. Laddha, and C. V-Gonzalez. Sensor fusion for joint 3D object detection and semantic segmentation. In *Proc. of the IEEE CVPRW*, 2019.

- [21] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *Proc. of the IEEE CVPR*, 2019.
- [22] S. Shi, X. Wang, and H. Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *Proc. of the IEEE CVPR*, 2019.
- [23] Y. Zhou, P. Sun, Y. Zhang, D. Anguelov, J. Gao, T. Ouyang, J. Guo, J. Ngiam, and V. Vasudevan. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds. In *Proc. of the CoRL*, 2019.
- [24] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia. STD: Sparse-to-dense 3D object detector for point cloud. In *Proc. of the IEEE ICCV*, 2019.
- [25] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, and N. Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *Proc. of the ICRA*. IEEE, 2019.
- [26] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff. Covernet: Multi-modal behavior prediction using trajectory sets. In *Proceedings of the IEEE CVPR*, 2020.
- [27] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proc. of the IEEE CVPR*, 2018.
- [28] R. O. Chavez-Garcia and O. Aycard. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 2015.
- [29] S. Chadwick, W. Maddetn, and P. Newman. Distant vehicle detection using radar and vision. In *Proc. of 2019 ICRA*. IEEE, 2019.
- [30] M. Meyer and G. Kuschk. Deep learning based 3d object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*. IEEE, 2019.
- [31] B. Major, D. Fontijne, A. Ansari, R. Teja Sukhavasi, R. Gowaikar, M. Hamilton, S. Lee, S. Grzechnik, and S. Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proc. of the IEEE ICCVW*, 2019.
- [32] L. Sless, B. El Shlomo, G. Cohen, and S. Oron. Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation. In *Proc. of the IEEE ICCVW*, 2019.
- [33] J. Lombacher, K. Laadt, M. Hahn, J. Dickmann, and C. Wöhler. Semantic radar grids. In *Proc. of the IEEE IV*. IEEE, 2017.
- [34] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection, 2020.
- [35] M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3D object detection. In *Proc. of the ECCV*, 2018.
- [36] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin. *Principles of modern radar*. Scitech Publishing, 2010.
- [37] M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of ECCV*, 2018.
- [38] M. Liang, B. Yang, W. Zeng, Y. Chen, R. Hu, S. Casas, and R. Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE CVPR*, 2020.
- [39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [40] G. P. Meyer and N. Thakurdesai. Learning an uncertainty-aware object detector for autonomous driving. *arXiv preprint arXiv:1910.11375*, 2019.
- [41] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. NuScenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

LiRaNet: End-to-End Trajectory Prediction using Spatio-Temporal Radar Fusion (Appendix)

Meet Shah^{*1}, Zhiling Huang^{*1}, Ankit Laddha^{*1}, Matthew Langford¹, Blake Barber¹, Sidney Zhang¹, Carlos Vallespi-Gonzalez¹, Raquel Urtasun^{1,2}

¹Uber Advanced Technologies Group, ² University of Toronto
`{meet.shah, zhiling, aladdha, mlangford, bbarber, sidney, cvallespi, urtasun}@uber.com`

In this appendix, we first outline implementation details for training LiRaNet. Then, we discuss the network architecture in detail. Next, we discuss additional results in terms of detection performance, and analysis based on different object characteristics. We then analyze the run-time impact of adding radar and present additional qualitative results.

1 Implementation Details

We first describe the implementation details of our proposed method, LiRaNet, on both datasets. On nuScenes, we discretize x and y axis with a resolution of 10 cells per meter and z axis with a resolution of 5 cells per meter. We use data augmentation to further increase the amount of training data on nuScenes. We interpolate labels from key frames to non-key frames to increase the number of training samples. We further increase the training data by randomly rotating, translating and scaling point clouds and labels. All past point clouds and future trajectory labels are augmented using the same random transformation. For each training sample, we use translation between $[-1, 1]m$ for the x and y axes, and $[-0.2, 0.2]m$ for the z axis. We use a rotation between $[-\pi/4, \pi/4]$ about the z axis and a scaling between $[0.95, 1.05]$ for all 3 axes.

On X17k, we discretize the x and y axis with a resolution of 6.4 cells per meter and z axis with a resolution of 5 cells per meter. This dataset contains labels for all the frames and has an order of magnitude more data than nuScenes. Thus, we do not perform any data augmentation.

Similar to previous work [1, 2], on nuScenes, we define our *vehicle* class to contain car, truck, bus, trailer and construction vehicles. On X17k, the same definition of the vehicle class is used. For both datasets, we only train and test on the objects which contain at least one lidar point. During training, we define a BEV cell as positive if it contains the center of the object. We use Adam optimizer and train with batch size of 32 on 16 GPUs for 2 epochs. We use 0.0004 as initial learning rate and decay it by 0.1 after 1.5 and 1.9 epochs.

2 Network Architecture and Object Detection

Figure 2 shows our network architecture for end-to-end trajectory prediction using lidar, radar and HD maps. Each sensor is pre-processed to generate multi-scale features. These features for all scales are concatenated and further processed using a backbone CNN to extract and combine multi-scale features using all the sensors. The output features of the backbone are used for detection using a full convolutional single shot detector. For each cell we predict a class probability and a bounding box using the multi-sensor features. We then pick top 200 cells using the predicted class probability and do NMS to remove duplicates. For each of the remaining detections (i.e. objects) we predict the trajectory using a crop of the same feature map used for detection. For each detected object, a large 40m x 40m rotated ROI centered around the object is cropped from the output feature volume produced by the backbone network. This crop is further passed through a CNN to generate final trajectory predictions.

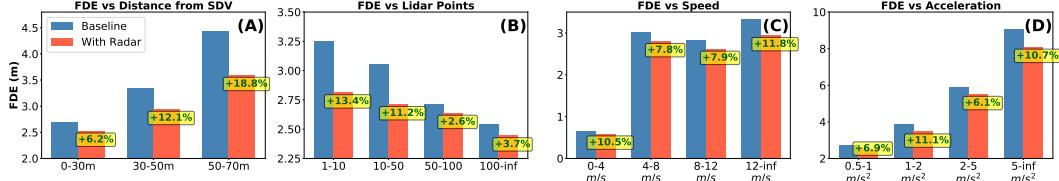


Figure 1: **Results by different object characteristics:** We show relative improvement (%) on nuScenes in FDE metric (m) by adding radar. (A) is binned by distance to AV, (B) is binned by number of lidar points, (C) is binned by the speed of the object, (D) is binned by the absolute value of acceleration of the object. These demonstrate that adding radar significantly improves over baselines in several challenging scenarios.

3 Comparison on Different Object Characteristics for nuScenes

Figure 1 shows improvement in trajectory prediction by adding radar along different factors which make trajectory prediction particularly challenging. Similar to X17k, we can observe that radar helps in all cases. It is especially beneficial in cases which are more challenging for lidar such as low lidar point density, rapidly changing dynamics due to acceleration, etc.

4 Detection Performance

In addition to the marked improvement on trajectory prediction as shown in the main paper, we have observed significant improvements in detection performance. We evaluate detection performance using the standard AP metric using IoU based association between ground truth and predicted bounding boxes. As we can observe from Table 1, addition of radar improves the detection performance on both datasets. We observe larger gains on nuScenes relative to X17k due to the relatively high sparsity of lidar points in nuScenes.

Method	Radar	nuScenes		X17k	
		AP _{0.5} ↑	AP _{0.7} ↑	AP _{0.5} ↑	AP _{0.7} ↑
MultiXNet* [3]	✗	71.54	58.47	84.02	74.13
LiRaNet (Ours)	✓	79.11	63.67	86.32	76.02

Table 1: **Improvement in detection performance by adding radar.** For nuScenes, we define our *vehicle* class to contain the car, truck, bus, trailer and construction vehicles. On X17k, the same definition of the vehicle class is used.

Method	Radar	Radar History (s)	Latency (ms)
MultiXNet* [3]	✗	-	38.02
LiRaNet (Ours)	✓	0.3s	38.51
LiRaNet (Ours)	✓	0.4s	39.96
LiRaNet (Ours)	✓	0.5s	40.98

Table 2: **Runtime Analysis on X17k.** All results are on a GeForce RTX 2080 Ti.

5 Runtime Analysis

Runtime latency is of prime importance for end-to-end trajectory prediction systems in order to deploy them in realtime applications.

The primary components of LiRaNet impacting runtime performance (relative to the lidar-only baselines) are: 1) the per cell feature extraction module and 2) the temporal fusion layers. We implement the per-cell feature extraction (including the k-nearest neighbour search) module as optimized CUDA [4] kernels. Since per cell feature extraction from radar points doesn't depend on data from other sensors, we pipeline it alongside with the map and lidar preprocessing modules to further reduce effective latency. The temporal fusion layers along with the core network backbone are further optimized for inference using TensorRT [5]. All results are computed on a GeForce RTX 2080 Ti GPU using batch size 1 and are averaged over 500 iterations. Latency of our network along with

the baselines can be found in Table 2. As we can see, overall impact of adding radar on runtime is relatively limited, $< 3ms$, even for the case of longer time horizon, $0.5s$ of radar history.

6 Qualitative Results

Figures 3, 4 and 5 show scenes from X17k where adding radar helps produce better trajectory predictions. We show lidar points with light blue, radar point velocities with orange, labels with white and model predictions with gold. The sensor data used in the model is depicted in the scene.

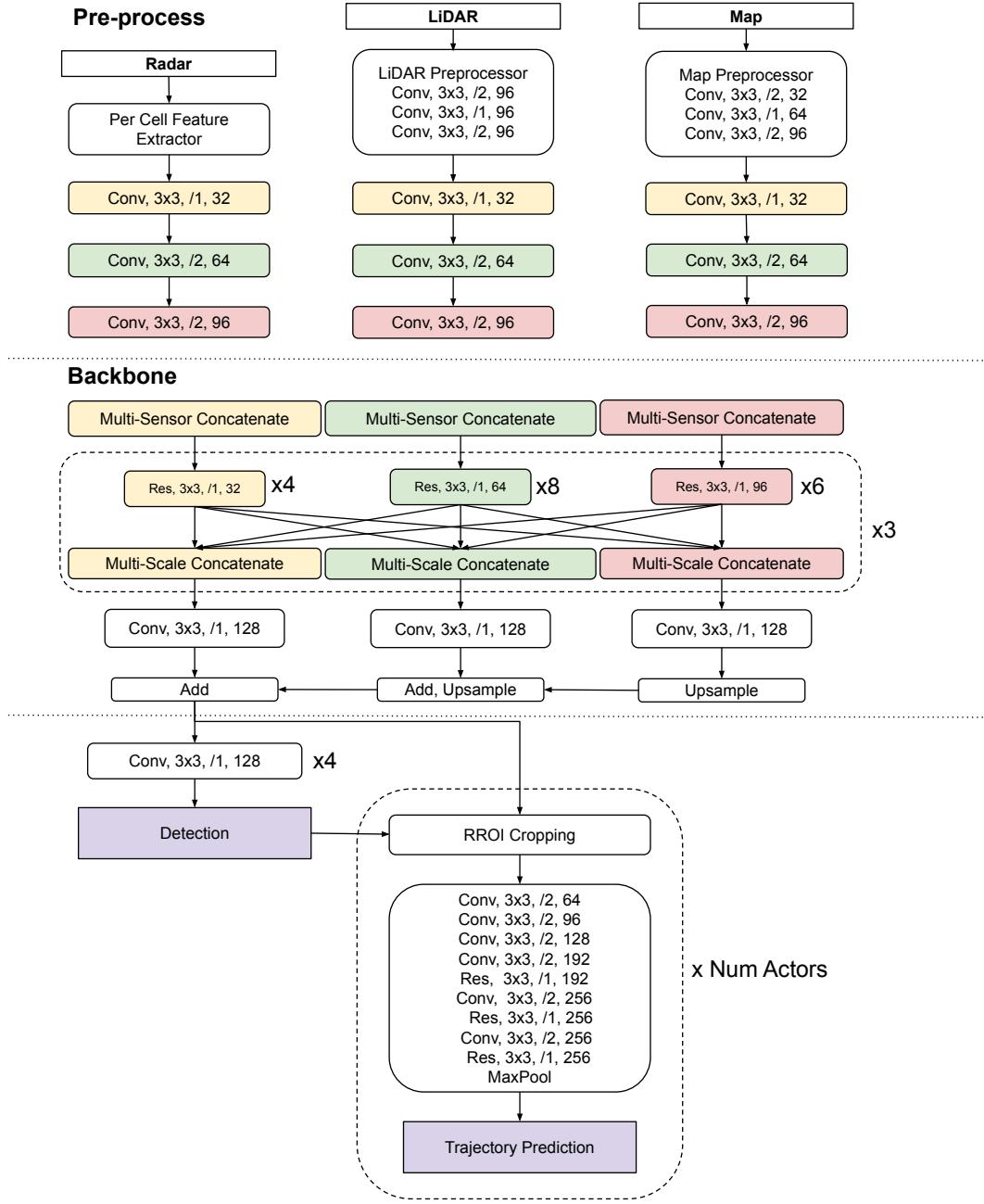


Figure 2: **Network Architecture:** This figure shows the detailed architecture of LiRaNet. Radar is fused with LiDAR and Map to generate multi-sensor feature volume in BEV which is used for detection and trajectory prediction. Each learnable operator is denoted as $(op, kernel, stride, number of features)$. *Conv* denotes convolution and *Res* denotes a residual block.

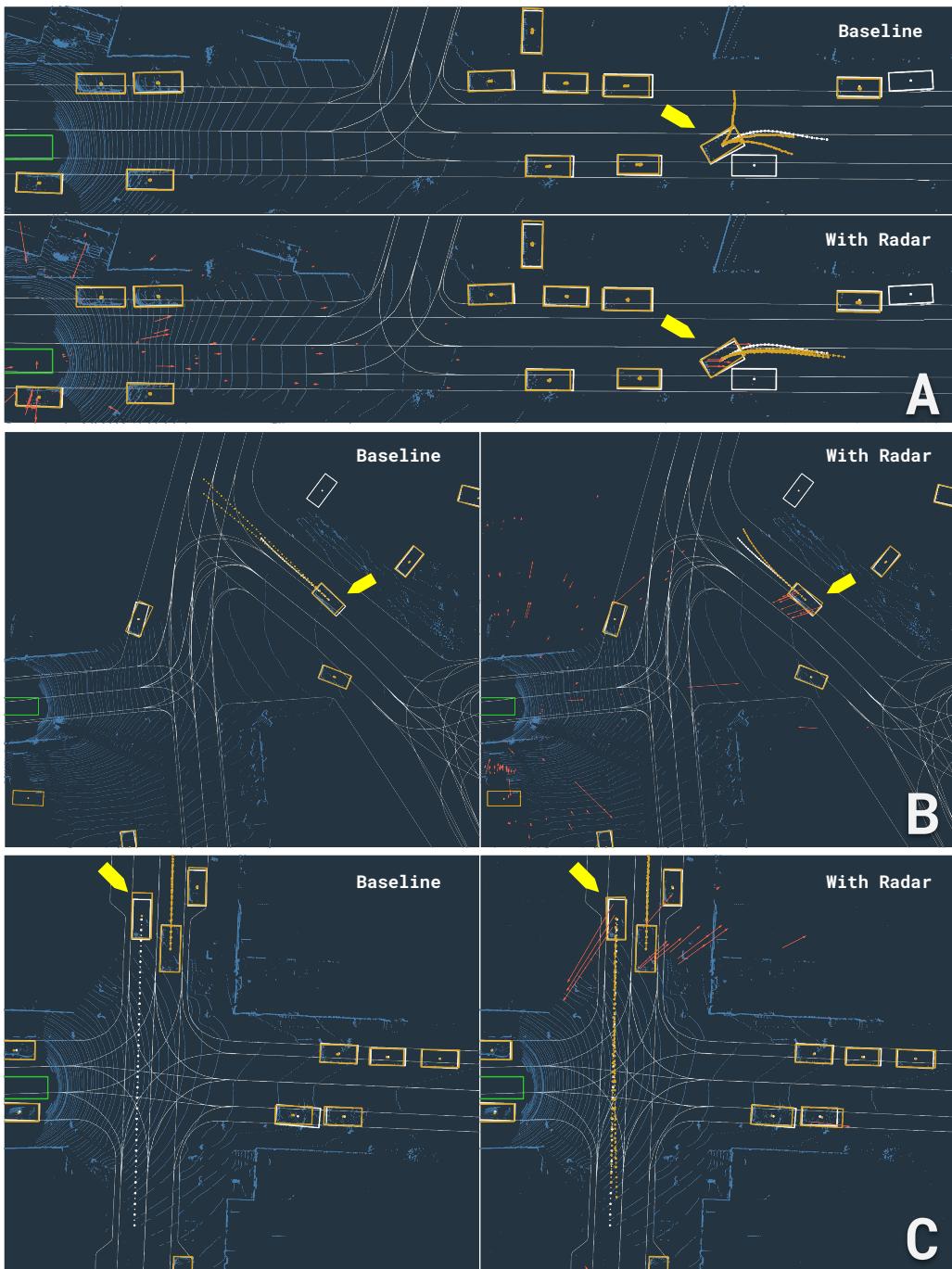


Figure 3: Improvements in Prediction. These three scenes show where radar improves trajectory prediction. The vehicle in **A** is accelerating into the street. Given the range and high acceleration of the vehicle, radar shows a clear benefit in estimating object trajectories with respect to the lidar only baseline. Similarly, **B** shows a case where the addition of radar helps accurately predict the motion of a vehicle slowing down for a turn. In **C** a vehicle has just come from an area of occlusion resulting in sparse historical lidar data. The radar data lets the model quickly estimate an accurate future trajectory.

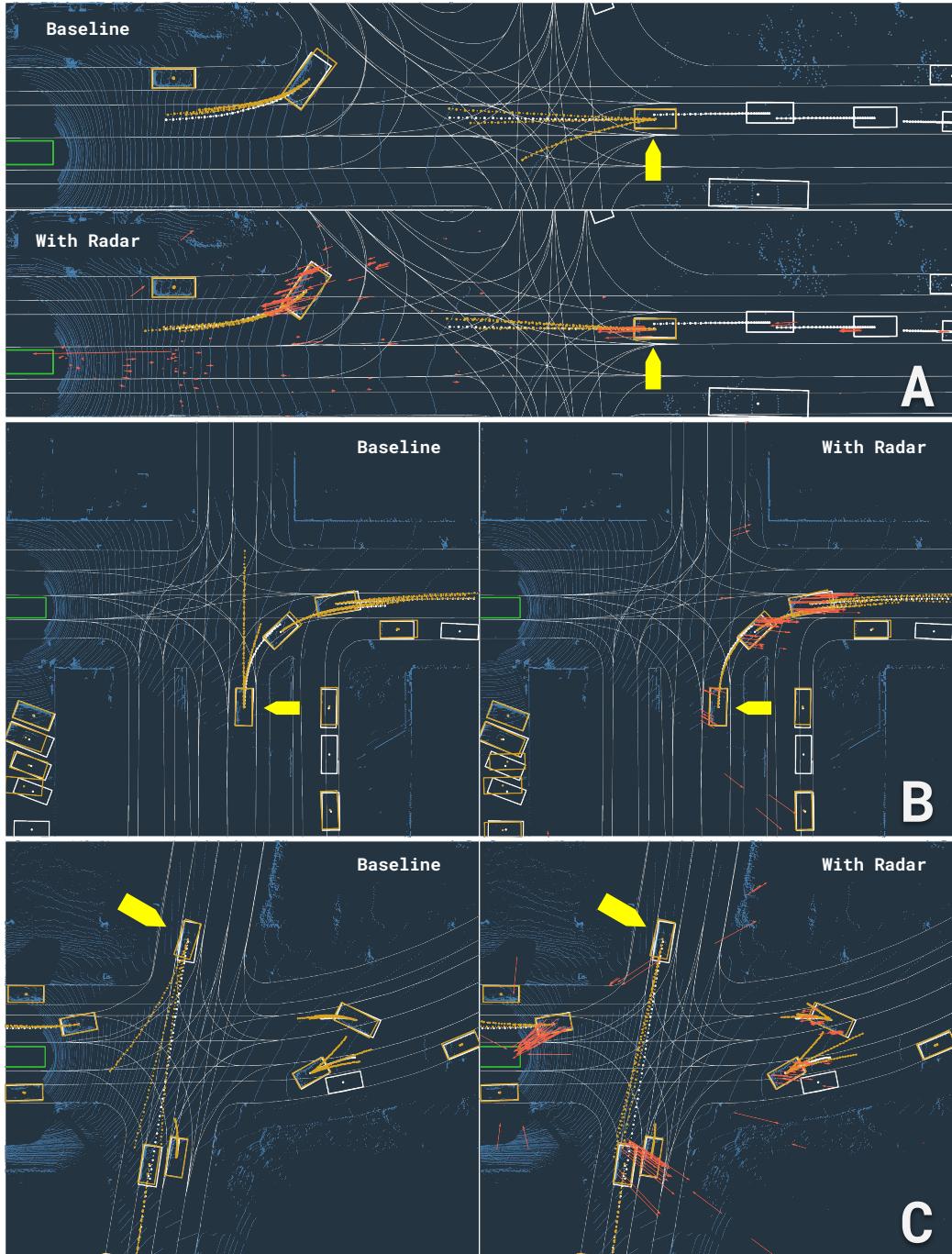


Figure 4: Improvements in Mode Ambiguity Resolution. Here we show three cases where radar helps resolve ambiguities in future trajectory prediction. We show the top 3 hypotheses. Without radar, **A** predicts a hypothesis that the vehicle can turn left in front of the SDV with high confidence. Without radar, the vehicle in **B** is predicted to cross SDV’s path. Mispredictions like these could lead to uncomfortable hard breaking. In **C**, the vehicle trajectory predictions converge when the velocity information from (even sparse) radar is incorporated. In all cases addition of radar leads to more accurate mode prediction.

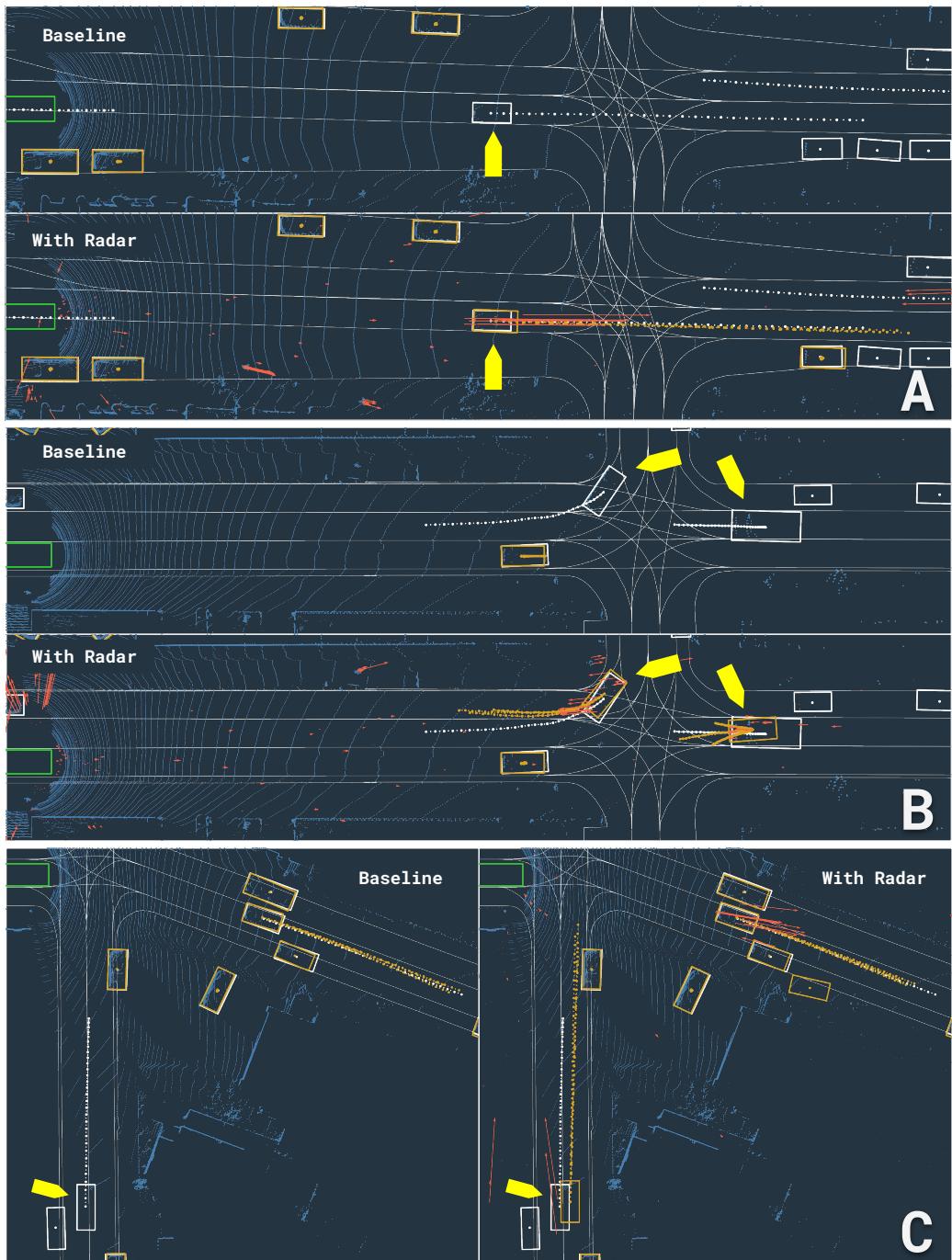


Figure 5: Improvements in Detection. These three scenes show where addition of radar improves the ability to detect and consequently predict trajectories of objects. In scene **A** the incline of the road and range of the object limit the number of lidar returns on the object. This results in less confident detections without radar. The two vehicles in **B** have been occluded by other objects in the scene. These objects predominantly detected by radar are noisier than if lidar was present. Similarly, in **C** the vehicle is emerging from occlusion. The number of lidar points on the object is too low to produce a detection without radar.

References

- [1] G. P. Meyer, J. Charland, S. Pandey, A. Laddha, C. V-Gonzalez, and C. K. Wellington. Laserflow: Efficient and probabilistic object detection and motion forecasting. *arXiv preprint arXiv:2003.05982*, 2020.
- [2] A. Laddha, S. Gautam, G. P. Meyer, and C. Vallespi-Gonzalez. Rv-fusenet: Range view based fusion of time-series lidar data for joint 3d object detection and motion forecasting. *arXiv preprint arXiv:2005.10863*, 2020.
- [3] N. Djuric, H. Cui, Z. Su, S. Wu, H. Wang, F.-C. Chou, L. S. Martin, S. Feng, R. Hu, Y. Xu, A. Dayan, S. Zhang, B. C. Becker, G. P. Meyer, C. V-Gonzalez, and C. K. Wellington. Multixnet: Multiclass multistage multimodal motion prediction, 2020.
- [4] NVIDIA, P. Vingelmann, and F. H. Fitzek. Cuda, release: 10.2.89, 2020. URL <https://developer.nvidia.com/cuda-toolkit>.
- [5] NVIDIA. Tensorrt, release: 7.1, 2020. URL <https://developer.nvidia.com/tensorrt>.