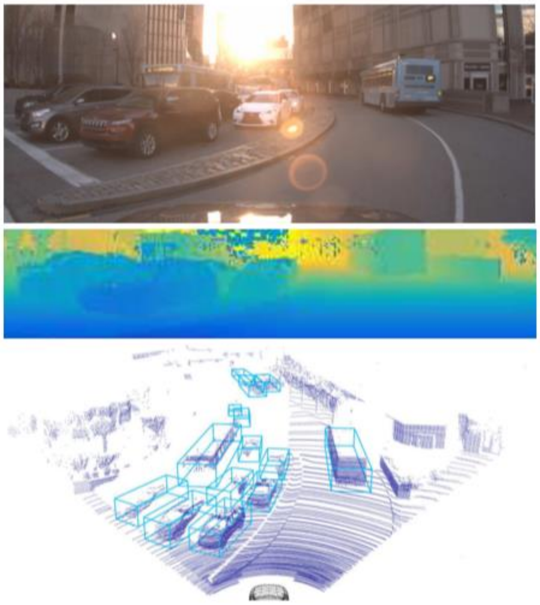
LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving: Analysis

- *by Yagna Hari Muni*

The paper discusses a study of an Efficient Probabilistic 3D Object Detector for Autonomous Driving. It combines qualitative and quantitative research to identify the most effective research solution. The paper includes about **30 citations**, the vast majority of which are about 3-D Object Detection. The authors proposed a new Range View approach to analyzing LiDAR data based on their findings and conclusions. Their method of modeling each detection as a distribution rather than a box improved overall performance. This document summarizes their research, objectives, analysis, forecasts, and conclusions.

A team of five authors from Uber Advanced Technologies Group, Gregory P. Meyer*, Ankit Laddha*, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K. Wellington, published this research paper in IEEE.

The authors clearly stated their goal of developing a computationally efficient method for processing LiDAR data in native range view. They model each detection as a distribution based on Multimodal distribution predictions in their approach. Even though the goal is clear, it would have sounded more appealing if they had included a quantitative analysis of their performance improvement or efficiency.



It can also be stated that the information in the abstract corresponds to the author's research paper. The abstract is straightforward and concise, with no information that is not covered in the paper.

Introduction:

The authors stated clearly in the introduction that LiDAR can be used for autonomous driving regardless of lighting conditions. They also stated that the purpose of this study is to improve a real-time solution for 3D object detection by investigating the uncertainties in existing approaches. They even compared LiDAR technology to camera image processing in terms of how it obtains a 3D output. This demonstrates how compelling their motivation for conducting the study is. Overall, the authors have effectively presented the uncertainties, methods used, and problem solution, but their thesis statement is unclear.

They presented that their approach was more efficient than existing models, first to capture the varying levels of uncertainty in order to make the driving system behave more efficiently around objects, but the thesis does not appear to be interesting due to the lack of quantitative parameters indicating how much accuracy they achieved compared to existing models.

Summary of Contributions:

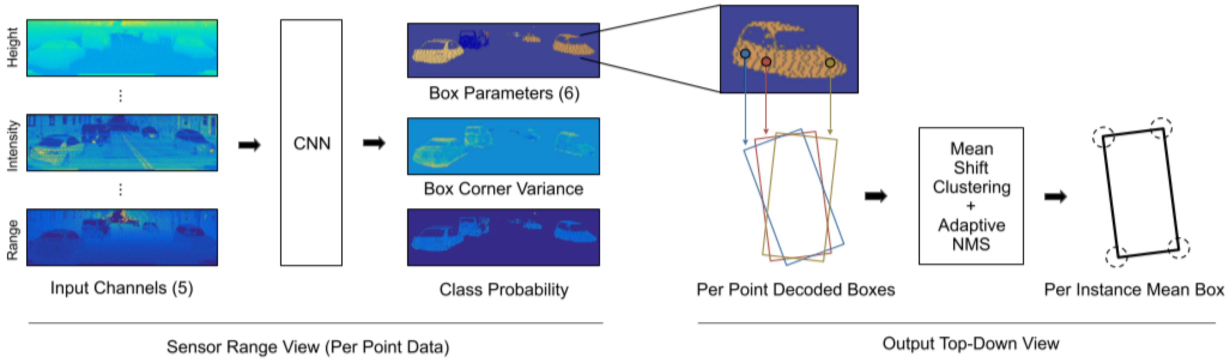
This study's main focus is on how well LiDAR can be used to improve autonomous driving capability. The paper focuses on how to process LiDAR data, what methods to use, and how the authors' approach contributed to higher accuracy.

The research was very interesting and original, being a real-time problem statement and the excitement that Elon Musk instilled in people about self-driving cars. In comparison to other published materials, this research paper proposes a better approach based on class probability and probability distributions for each detection box. The authors even claim that their method is the first to capture the uncertainties associated with overcoming occlusions through distribution modeling.

If the authors had defined the semantic uncertainty, localization uncertainty, epistemic uncertainty, and aleatoric uncertainty in relation to Probabilistic Object Detection, the user would have a clear picture.

Describe the proposed method:

The proposed method is well written on how input is perceived to make predictions to calculate the four corners of the bounding box, which means that the object is characterized into individual boxes.



The authors clearly stated the functionality of LiDAR and how it can generate the cylindrical image using its 64 Lasers. They use sensor data such as range, reflectance, azimuth, and laser id to create a five-channel input image for the network.

Once generated, the image contains a variety of objects at varying distances. The objects in the image can range from several thousand to a single point. They use deep layer aggregation network architecture to extract and combine some multi-scale features for this. The network architecture uses operations such as down sampling and up sampling to create a convolutional layer, which is then used to transform the resulting feature map to encoded predictions.

They stated that their network has been trained to predict probability distributions using a mixture model in order to overcome the challenges of multimodal when data is sparse. To characterize bounding boxes in this case of autonomous driving, they consider objects that are on the same ground plane. Instead of directly regressing the four corners of the box, they used a new approach to calculate the four corners of the bounding box by identifying the absolute center.

Because each object can contain multiple points, each point will predict the probability distribution of bounding boxes. Though these all point to the same object and should produce a similar distribution, there will be some noise. This noise is reduced by using mean shift clustering over box centers, which reduces the problem's dimensionality.

The authors used multi-class cross entropy loss to learn the class probabilities. They used focal loss, a modified version of cross entropy, to deal with class imbalance. They calculated classification loss for the entire image for this purpose. By referring to the research paper "Alex Kendall and Yarin Gal. "What uncertainties do we need in Bayesian deep learning for computer vision?", the total loss is calculated as the sum of the classification and regression losses.

Finally, non-maximum suppression (NMS) was used to remove the redundant bounding boxes. It can also be used to determine whether encapsulated objects are unique or to estimate the worst-case overlap of objects.

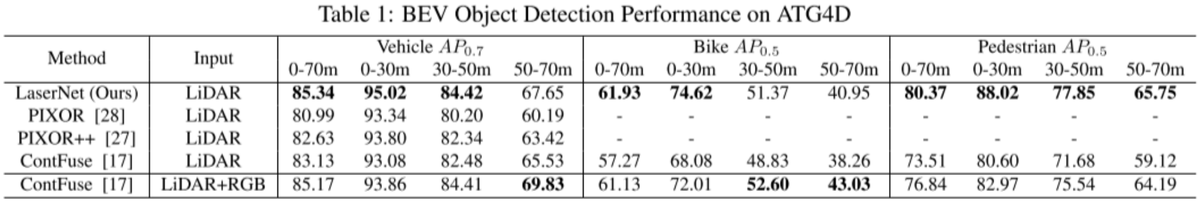


Table 1: BEV Object Detection Performance on ATG4D

Method	Input	Vehicle $AP_{0.7}$				Bike $AP_{0.5}$				Pedestrian $AP_{0.5}$			
		0-70m	0-30m	30-50m	50-70m	0-70m	0-30m	30-50m	50-70m	0-70m	0-30m	30-50m	50-70m
LaserNet (Ours)	LiDAR	**85.34**	**95.02**	**84.42**	67.65	**61.93**	**74.62**	51.37	40.95	**80.37**	**88.02**	**77.85**	**65.75**
PIXOR [28]	LiDAR	80.99	93.34	80.20	60.19	-	-	-	-	-	-	-	-
PIXOR++ [27]	LiDAR	82.63	93.80	82.34	63.42	-	-	-	-	-	-	-	-
ContFuse [17]	LiDAR	83.13	93.08	82.48	65.53	57.27	68.08	48.83	38.26	73.51	80.60	71.68	59.12
ContFuse [17]	LiDAR+RGB	85.17	93.86	84.41	**69.83**	61.13	72.01	**52.60**	**43.03**	76.84	82.97	75.54	64.19

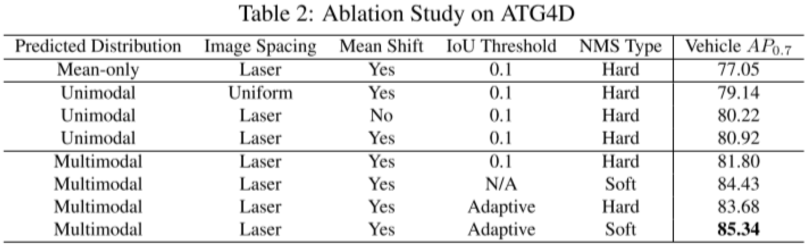


Table 2: Ablation Study on ATG4D

Predicted Distribution	Image Spacing	Mean Shift	IoU Threshold	NMS Type	Vehicle $AP_{0.7}$
Mean-only	Laser	Yes	0.1	Hard	77.05
Unimodal	Uniform	Yes	0.1	Hard	79.14
Unimodal	Laser	No	0.1	Hard	80.22
Unimodal	Laser	Yes	0.1	Hard	80.92
Multimodal	Laser	Yes	0.1	Hard	81.80
Multimodal	Laser	Yes	N/A	Soft	84.43
Multimodal	Laser	Yes	Adaptive	Hard	83.68
Multimodal	Laser	Yes	Adaptive	Soft	**85.34**

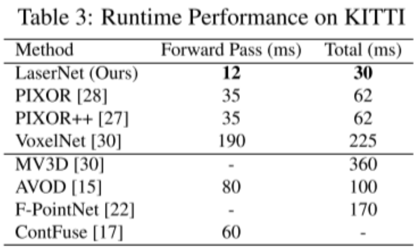


Table 3: Runtime Performance on KITTI

Method	Forward Pass (ms)	Total (ms)
LaserNet (Ours)	**12**	**30**
PIXOR [28]	35	62
PIXOR++ [27]	35	62
VoxelNet [30]	190	225
MV3D [30]	-	360
AVOD [15]	80	100
F-PointNet [22]	-	170
ContFuse [17]	60	-

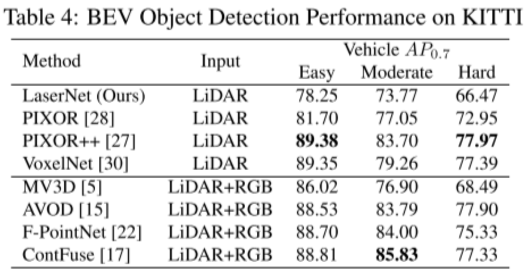
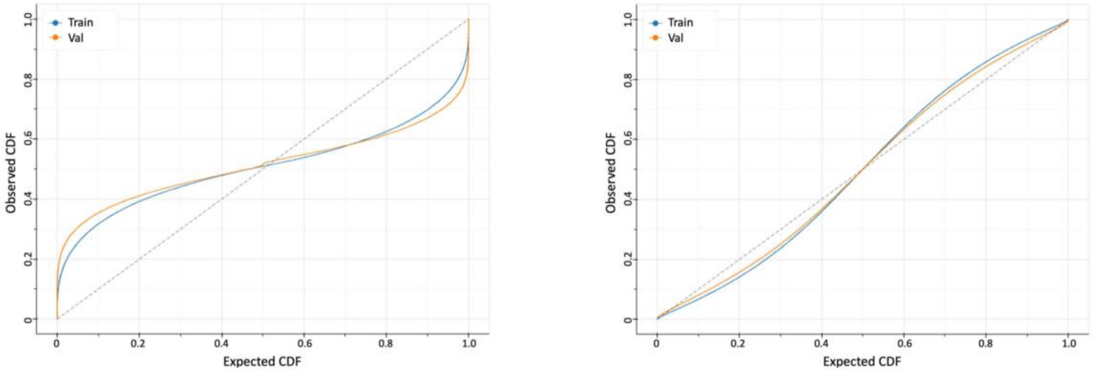


Table 4: BEV Object Detection Performance on KITTI

Method	Input	Vehicle $AP_{0.7}$		
		Easy	Moderate	Hard
LaserNet (Ours)	LiDAR	78.25	73.77	66.47
PIXOR [28]	LiDAR	81.70	77.05	72.95
PIXOR++ [27]	LiDAR	**89.38**	83.70	**77.97**
VoxelNet [30]	LiDAR	89.35	79.26	77.39
MV3D [5]	LiDAR+RGB	86.02	76.90	68.49
AVOD [15]	LiDAR+RGB	88.53	83.79	77.90
F-PointNet [22]	LiDAR+RGB	88.70	84.00	75.33
ContFuse [17]	LiDAR+RGB	88.81	**85.83**	77.33

Experimentation:

To put their proposed approach to the test, the authors evaluated and compared it on two data sets.



(a) Calibration on KITTI

(b) Calibration on ATG4D

1. **Large Scale ATG4D:** Using ATG4D evaluation metrics (5000 sequences for training and 500 for validation), the team was able to achieve state-of-the-art performance in the 0-70 meter range. Not only did it outperform the LiDAR only method at all ranges, but it also outperformed the LiDAR + RGB method on vehicles and bikes at long rates. The authors claim that because their approach does not discretize the input data into voxels, it can better identify the pedestrians.

   Examining the various aspects of the proposed method, the authors claim that predicting a distribution of bounding boxes rather than the mean is the most significant improvement. Non-uniformly spaced lasers in the Velodyne 64E LiDAR helped in mapping points to rows based on laser id and processing data as the sensor captures it directly. Mean shift clustering improved performance even more by reducing noise. Finally, by using the adaptive NMS algorithm, the probabilistic interpretation is preserved, and better performance is achieved.

   According to the authors, the proposed method is twice as fast as the state-of-the-art method measured on an NVDIA 10080Ti GPU. The shorter runtime is achieved by operating on a small dense range view image rather than a sparse bird's eye view, which must be considered.

2. **Small Scale KITTI:** KITTI's evaluation metrics include approximately 7481 training sweeps and 7518 testing sweeps captured by Velodyne 64E LiDAR. Because the data set was small in comparison to ATG4D, the authors' proposed approach underperformed when compared to the state-of-the-art method.

The team plotted the CDF between expected and actual distributions to assess the quality of the predicted distributions. This evaluation is carried out on both the KITTI and the ATG4D datasets. The model is trained on both datasets to predict a unimodal probability distribution. The model is unable to correctly estimate the probability distribution using the small scale KITTI dataset. The model, on the other hand, is capable of learning the distribution precisely on the large scale ATG4D dataset. They hypothesize that learning the distribution necessitates the network seeing many more examples than are available in the KITTI training set, which explains the difference in model performance between these two datasets.

Strength and Weaknesses

Strengths:

1. Estimating probability distributions over box detections rather than mean and variance as in previous approaches.
2. Comparison of quantitative performance analysis with state-of-the-art existing methods.
3. Key metrics for comparing large-scale ATG4D data to small-scale KITTI data
4. Citations that explain the existing methods in detail
5. Non-maximum suppression (NMS) is used to eliminate redundant bounding boxes.

Weaknesses

1. Low performance on KITTI data with a small sample size
2. Inconsistencies in explaining or describing the methods cited in other research papers. For example, there is insufficient information about semantic and localization uncertainty.
3. Inadequate conclusion that does not describe the methods used to improve accuracy. Predicting a distribution over a set of bounding boxes, for example, loss analysis, or mean shift clustering.

Things that can be improved:

1. Describe the methods cited in other research papers that you considered. For example, what are CDF plots, semantic and localization uncertainty?
2. In the abstract and conclusions, quantitative analysis can be decribed.
3. By comparing more datasets, you can obtain evaluation metrics.

The paper is well written and covers all of the existing approaches as well as the improvements made by the probabilistic approach to improve accuracy. It was clearly described which techniques aided in improving the accuracy of that factor. Analyzing the paper as a whole, the text is fairly clear and easy to understand, despite a few omissions due to undescribed references and short forms.

Conclusion

Detectors that operate in a bird's eye view or directly on a 3D point cloud were previously preferred by the research community due to their advantages. However, the range view presents significant challenges due to its varying scale, object shape, and tasks to handle occlusions. Using the authors' approach and a large dataset, the authors propose that it is possible to overcome these challenges and produce competitive results.

However, this cannot be achieved on a smaller dataset due to operations in the sensor's native view and predicting probability distribution over bounding boxes.