High Dimensional Frustum PointNet for 3D Object Detection from Camera, LiDAR, and Radar

Leichen Wang[#1,2], Tianbai Chen[#1,3], Carsten Anklam[#1] and Bastian Goldluecke[#2]

[#1]Daimler AG, [#2]University Konstanz, [#3]KIT, Germany

leichen.wang, tianbai.chen, carsten.anklam@daimler.com

bastian.goldluecke@uni-konstanz.de

Overview

I. Motivation :Why sensor fusion?

II. Baseline: Frustum PointNet

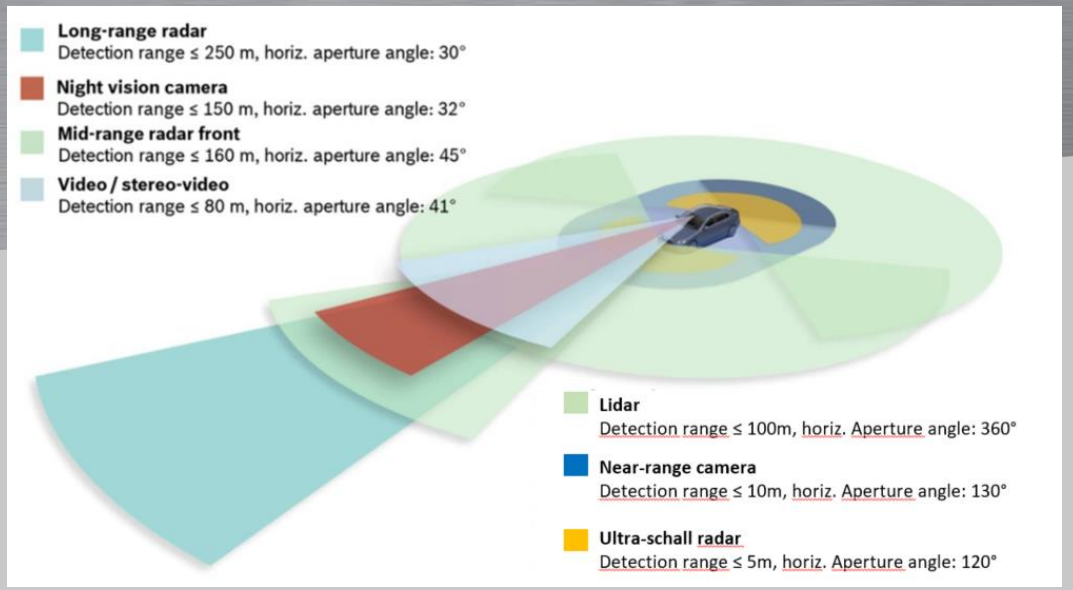
III. High Dimensional Frustum PointNet

IV. Experimental steup

V. Conclusions

Motivation :Why sensor fusion?

- Single-sensor dependence
  - Dangerous
  - Lack of redundancy
- Exploit positive sensor attributes
  - Complimentary sensor technologies
  - Enable higher performance



Sensor	Range	Resolution	Non-metal Object	Stationary Object	All weather
MMR	😐	☹	☹	😐	☺
LLR	☺	☹	☹	☹	☺
Camera	☹	☺	☺	☺	☹
LiDAR	😐	☺	☺	☺	☹

Motivation :Why sensor fusion?

Ground Truth	Radar	Camera	Lidar	Sensor Fusion
Car @150m		Noise	Noise	Car
Motorcycle @100m				Motorcycle
Bike rider @50m				Bike rider
Tire@100m	Noise		Obstacle	Brake or ignore?

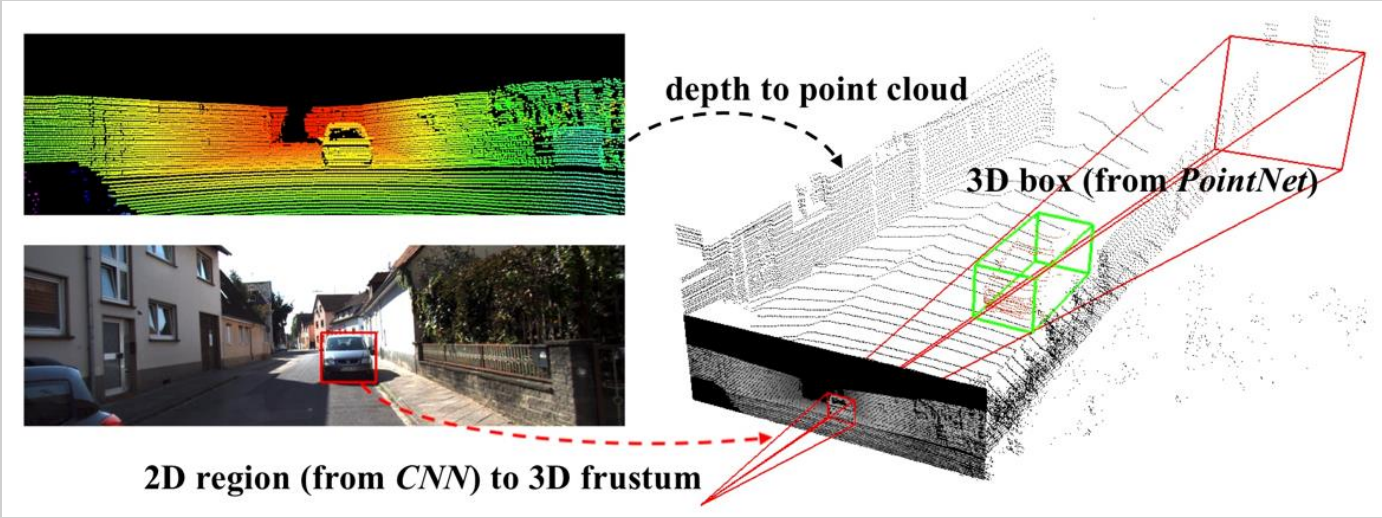
- Object Classifciation in range & FOV of interest must be comparable.
- Sensor fusion can help auto understand the road and traffic environment better.

Baseline: Frustum PointNet*

1. They first generate 2D object region proposals in the RGB image using a CNN.

2. Each 2D region is then extruded to a 3D viewing frustum.

3. Finally, frustum PointNet predicts a 3D Bbox for the object from the points in frustum.

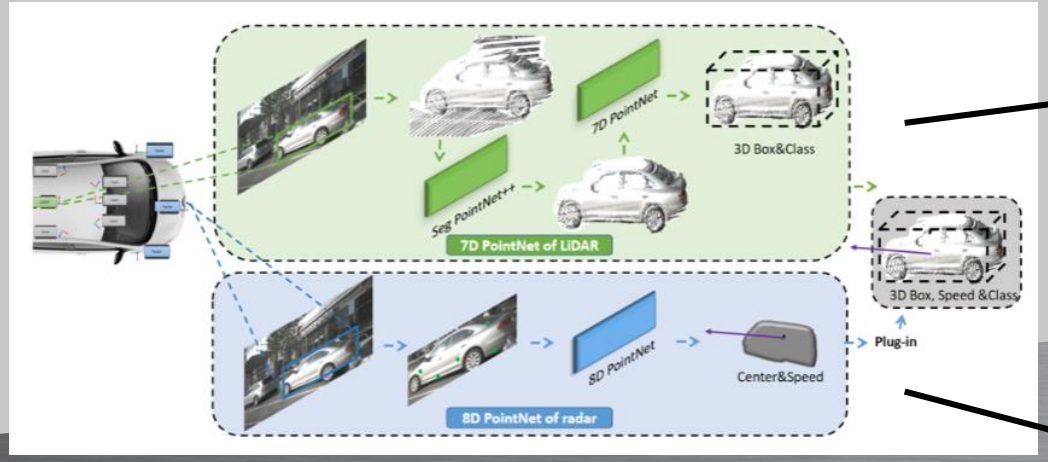


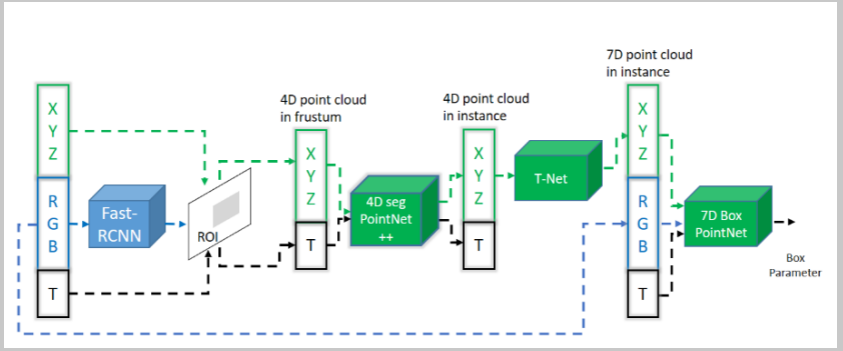
*Frustum PointNets for 3D Object Detection from RGB-D Data (CVPR 2018)

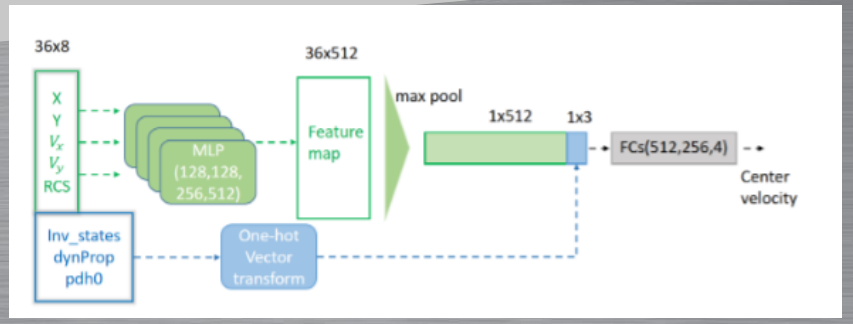
High Dimensional Frustum PointNet

- Our framework consists of three modules: 2D frustum proposal from camera, 7D PointNet for LiDAR pointclouds, and 8D PointNet for radar point clouds.

- In the end, all features and modules are combined and refined to estimate a 3D bounding box, which parameterized by its class, width, length, height, center X, Y, Z, heading angle and velocity.







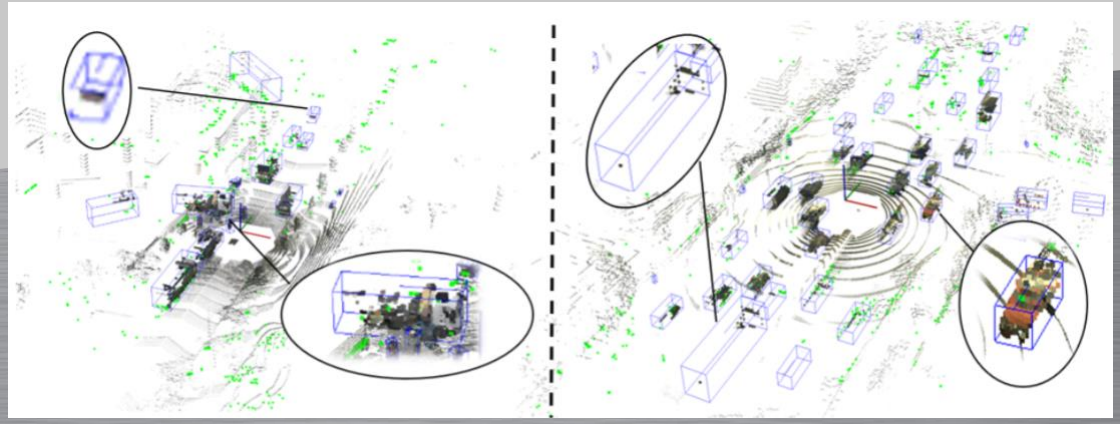
Experimental steup

1. We evaluate our method on the nuScenes benchmark for 3D object detection.

2. Full sensor suite (1x LIDAR, 5x RADAR, 6x camera, IMU, GPS)
   I. 1000 scenes of 20s each
   II. 1,400,000 camera images
   III. 390,000 LiDAR sweeps (Velodyne 32E)
   IV. 200,000 Radar sweeps (Continental ARS40X)
   V. 1.4M 3D bounding boxes manually annotated for 23 object classes

3. nuScenes detection score (NDS): consolidate the metrics by computing a weighted sum:
   mAP, mATE, mASE, mAOE, mAVE and mAAE.
   Average Translation Error (ATE)
   Average Scale Error (ASE)
   Average Orientation Error (AOE)
   Average Velocity Error (AVE)
   Average Attribute Error (AAE)

$$NDS = 0.1\left[5\,mAP + \sum_{mTP \in \mathbb{TP}} (1 - \min(1, mTP))\right]$$



Experimental steup

	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
MEGVII[43]	52.8	0.300	0.247	0.380	0.245	0.140	63.3
SARPNET[40]	32.4	0.400	0.249	0.763	0.272	0.090	48.4
PointPillars[15]	30.5	0.517	0.290	0.500	0.316	0.368	45.3
WYSIWYG[12]	35.0	0.382	0.245	0.554	1.000	0.379	41.9
MAIR[34]	30.4	0.738	0.263	0.546	1.553	0.134	38.4
F-PointNet[27]	32.5	0.544	0.433	0.781	0.518	0.079	42.7
Ours	36.6	0.483	0.397	0.670	0.481	0.091	46.8

Fig. 10. Comparison with state-of-the-art methods and baseline on the nuScenes benchmark test set.

	mean	Car	Ped	Bicycle	Bus	Barrier	TC	Truck	Trailer	Moto	Cons.Veh
MEGVII[43]	**52.8**	**81**	**80**	22	**55**	**65**	**71**	**49**	**53**	**51**	11
SARPNET[40]	32.4	60	70	14	19	38	45	19	18	30	12
PointPillar[15]	30.5	70	60	2	34	33	30	25	17	20	4
MAIR[34]	30.4	48	37	24	19	51	49	49	33	18	7
F-PointNet[27]	32.5	44	54	27	24	43	46	23	11	41	11
Ours (test set)	36.6	48	56	**28**	33	44	48	29	18	41	**23**
Ours (eva set)	46.8	45	70	46	34	49	70	39	13	51	41

Fig. 11. mAP for different categories compared to state-of-the-art methods. In evaluation set (eva set), 2D ground truth is given.

Experimental steup

- In comparison with Frustum PointNet, PointPillar and SARPNet, our approach achieves better performance on most of the categories, which proves the importance and improvement of high dimensional feature learning for 3D object detection.

- While MEGVII expands the original dataset by 4.5 times (from 28130 to 128100 samples); training with augmented data takes almost five times longer than in our framework.

- Thus, our method achieves a better balance with regard to training time and accuracy. Despite this, the performance of our network on unbalanced classes e.g. bicycle and construction vehicle significantly exceeds in mAP as compared to the other methods.

Conclusions

- To the best of our knowledge, our work is the first deep learning based approach for ”all-sensors” 3D object detection within a real-world scene.
- We propose a novel method termed High Dimensional Frustum PointNet for 3D object detection from raw data of cameras, radars, and LiDARs.
- From a practical perspective, we design a plug-in framework to extract features from radar point clouds efficiently.
- We perform extensive experiments and show that our approach achieves competitive or better results than current state-of-the-art methods for object detection.

Conclusions

- Through statistics and analysis, we found that, there are only 18% of pedestrians,11% bicycles and 43% of cars in our val set have valid radar points. From the radar specifications, we know that Continental ARS40X is mainly designed for industrial use cases instead of traffic scenery
- For future research, we intend to resample both LiDAR and radar point clouds into voxels, which might further improve the detection performance as well as run-time of our network.
- We intend to explore how to create high-quality feature maps from automotive radar instead of industrial radar.
- Finally, we will explore the detection performance in the range of 50~150 meters.

DAIMLER

Thanks for your attention!

leichen.wang, tianbai.chen, carsten.anklam@daimler.com
bastian.goldluecke@uni-konstanz.de