Radar and Camera Early Fusion for Vehicle Detection in ADAS

Teck-Yian Lim[1], Amin Ansari[2], Bence Major[3], Daniel Fontijne[3], Michael Hamilton[2], Radhika Gowakar[2], Sundar Subramanian[2]

[1]University of Illinois at Urbana-Champaign, [2]Qualcomm Technologies Inc., [3]Qualcomm AI Research

Motivation

- Traditional radar processing algorithms discard significant contextual information
- Current use of radar in ADAS and autonomous driving systems relies on sparse points obtained from traditional radar processing pipelines
- FMCW radars are already present in modern ADAS and are much more affordable compared to LiDAR systems

Introduction

We introduce **FusionNet**, a network architecture and training scheme for ingesting multiple dense sensor inputs.

1. Initial feature extraction in sensor space
2. Spatial transform to common space
3. Outperforms independent sensors
4. Proposed training scheme to handle different convergence rates

Dataset

Description:
150k automatically annotated frames of San Diego highways. 5000 frames were set aside as the validation set.

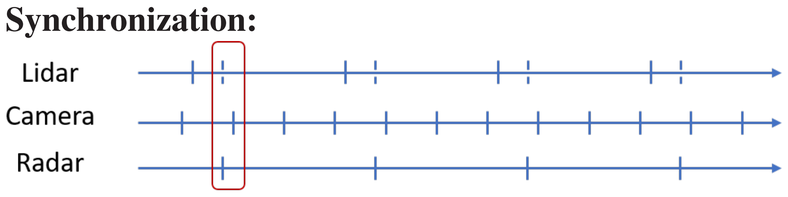
Sensors:

- Radar: Raw 77GHz FMCW, 180° FOV@125fps
- Camera: 150°FOV 1280×800@30fps
- LiDAR: Velodyne 32-line@10Hz

Radar Data Description: Our radar input differs significantly from existing public datasets, radar data in these datasets are sparse point clouds with velocity. Instead, we utilize minimally processed radar range-azimuth signals.

Training data generation: We generated ground truth bounding boxes using LiDAR point clouds, followed by manual inspections of the generated boxes. Different parameters may be utilized for difference sequences, and non-causal processing and filtering were also used to improve bounding box quality.

Synchronization:



Radar/Camera pair with interpolated bounding box from Lidar

We synchronized our sensors to match the lowest rate sensor on our platform (i.e. 10Hz LiDAR). A training tuple consists of a radar frame as timestamp reference, together with the nearest camera frame. The corresponding ground truth boxes are interpolated from the LiDAR pipeline.

Future Work

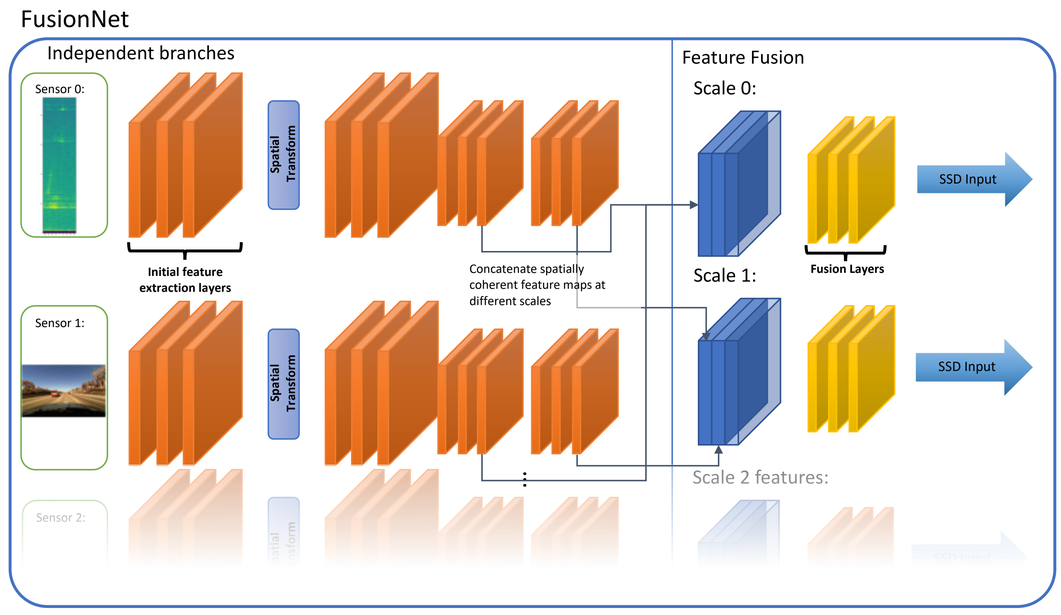
- Incorporate manual annotation in the dataset
- Utilize multiple frames/temporal information
- Utilize velocity information from radar
- Greater variety of frames and scenarios (*e.g.* urban scenes, pedestrians)
- Include more sensors and branches for a holistic 360° perception around the vehicle

Website Link

Please visit our project website for more examples and details.

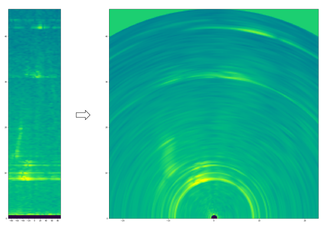
Our Approach

High-Level Overview

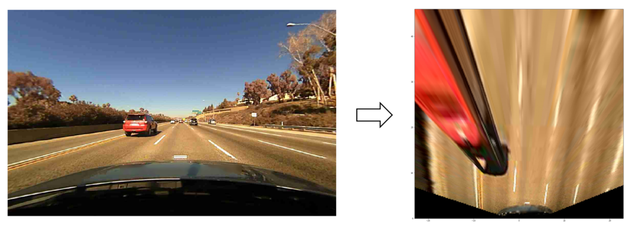


Branch Details: Spatial transforms

Radar Branch



Camera Branch



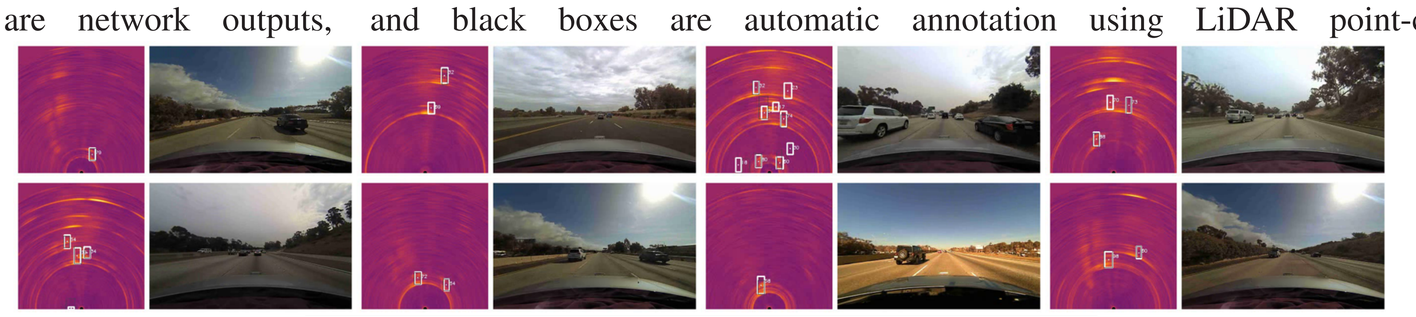
Fusion Layers: Output of branches are spatially coherent; therefore, we can now combine the output of the branches and use additional convolutional layers to combine features.

Detection Outputs: Detection approach is SSD inspired; however, since we are performing prediction in real-world metric space, object sizes of specific classes do not defer significantly as compared to prediction in image space. As a result, we do not require multiple scales to detect objects of the same class, but multiple scales are required for detection of objects classes of different sizes.

Dealing with different convergence rates for branches: Training is done in stages, branch with the slower convergence rate is trained to convergence, weights are then frozen, and additional branches are added and trained to convergence.

Results

We observe that our proposed architecture outperforms each of the individual sensors working independently, combining the advantage of each sensor. In the samples below, white boxes are network outputs, and black boxes are automatic annotation using LiDAR point-clouds.



	Fusion (SGD)	Fusion (ADAM)	Camera Only	Radar Only
mAP	**73.5%**	71.7%	64.65%	73.45%
Position x	**0.145m**	0.152m	0.156m	0.170m
Position y	**0.331m**	0.344m	0.386m	0.390m
Size x	0.268m	0.261m	**0.254m**	0.280m
Size y	**0.597m**	0.593m	0.627m	0.639m
Matches	**8695**	8597	7805	8549

Ablation and Sensitivity

We evaluated contributions of individual branches by selectively zeroing out/adding noise to inputs. We observed that the loss or degradation of each input results in a significant drop in the network performance.

	Fusion	Camera 0	Radar 0	Camera + Noise	Radar + Noise
mAP	73.5%	55.0%	19.4%	61.2%	71.9%
Position x	0.1458m	0.1883m	0.1816m	0.1667m	0.1524m
Position y	0.3315m	0.4297m	0.3602m	0.3847m	0.3360m
Size x	0.2688m	0.3042m	0.3230m	0.4126m	0.2686m
Size y	0.5975m	0.7829m	0.5653m	0.7022m	0.5853m
Matches	8695	8259	3051	8004	8554