DeepBEV: A Conditional Adversarial Network for Bird's Eye View Generation

Helmi Fraser
hmf30@hw.ac.uk

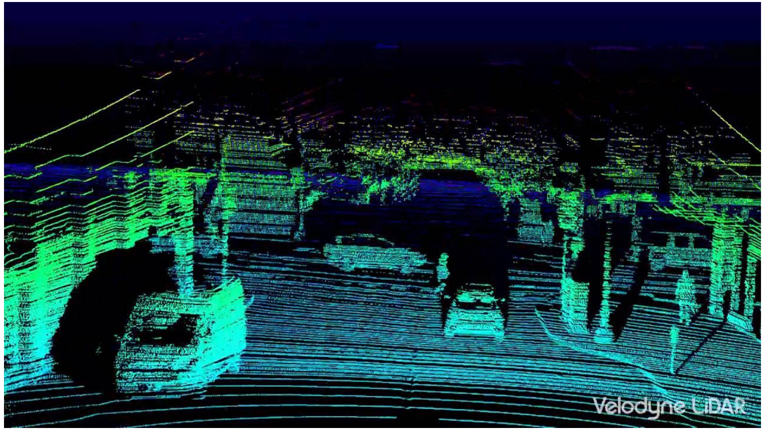
Sen Wang
s.wang@hw.ac.uk

Heriot-Watt University
Edinburgh Centre for Robotics

Motivation

It is vitally important that an autonomous vehicle perceives its environment

These sensors are excellent for 3D perception, but they are expensive

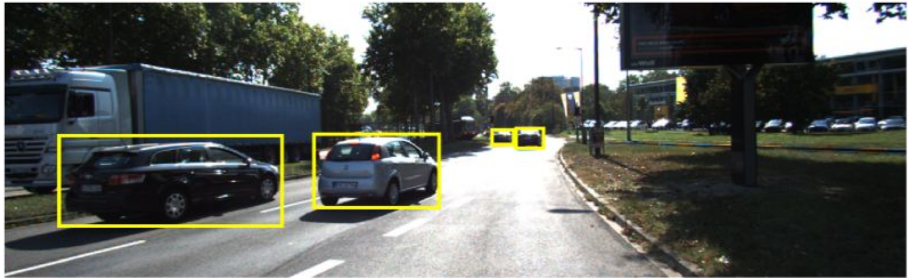


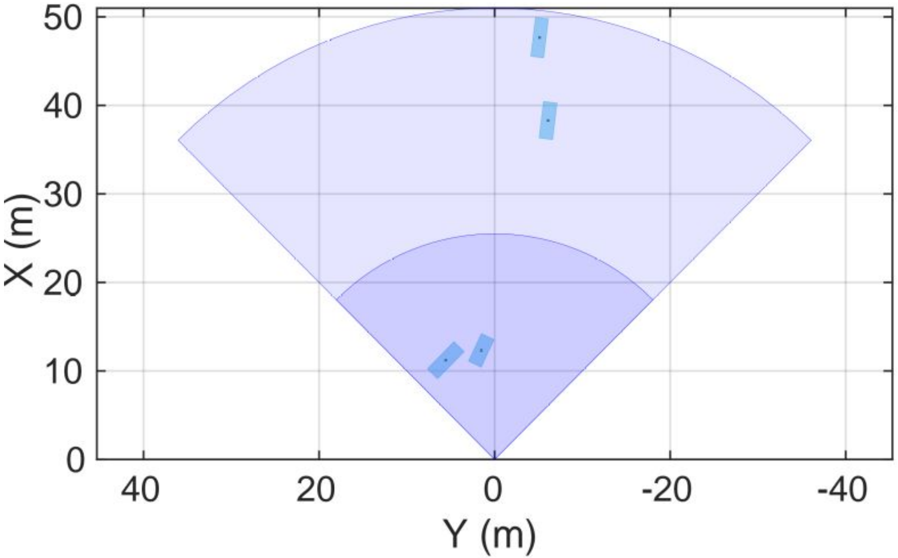
Courtesy of Velodyne

Courtesy of Navtech

Motivation

Efficient and interpretable representation of *semantically significant* objects

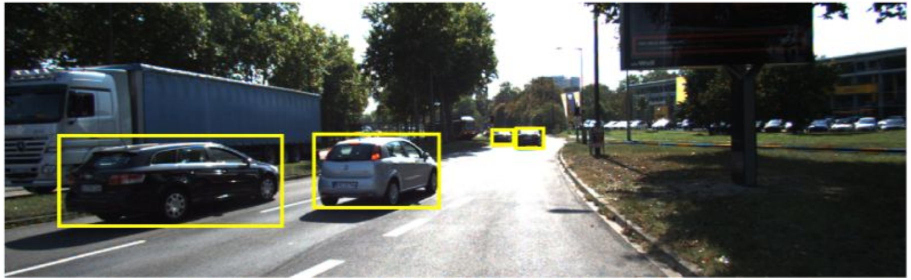


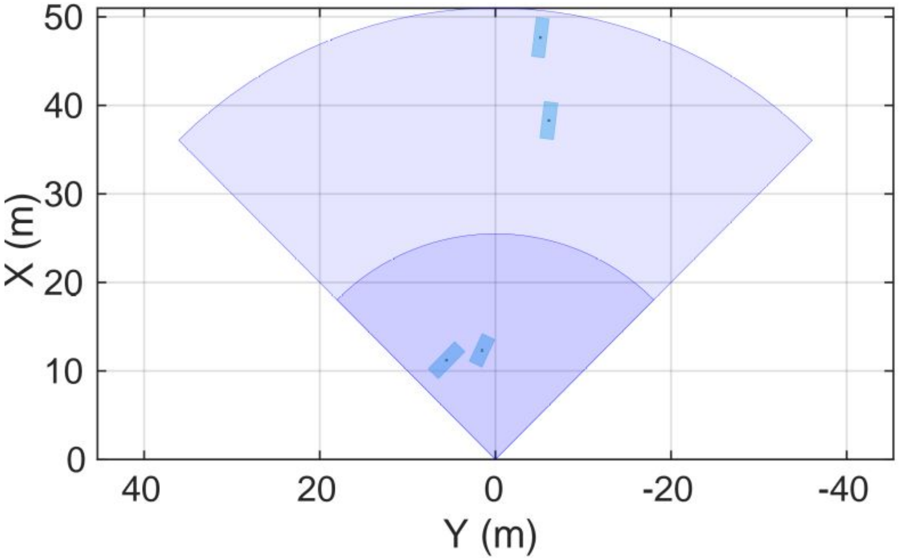


Solution
Bird's Eye View

Benefits:

- Low memory requirement
- Simple
- Only requires a camera

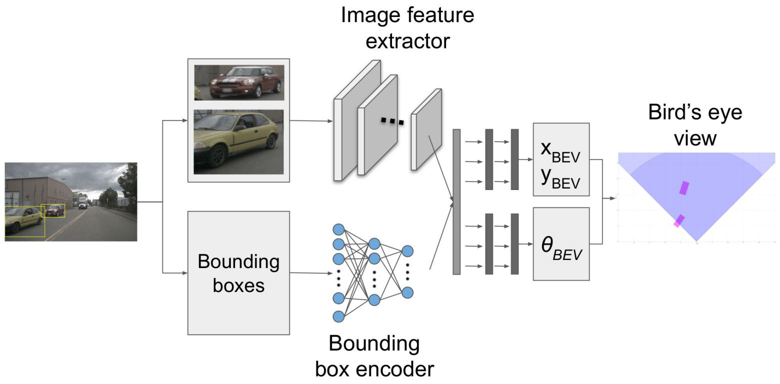




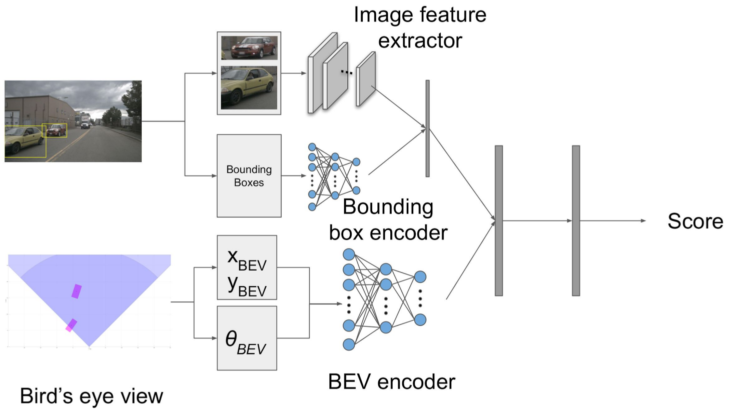
Method

Formulate as an adversarial learning task:

- A generator outputs BEV representations
- A critic scores this representation



Generator



Critic

Method

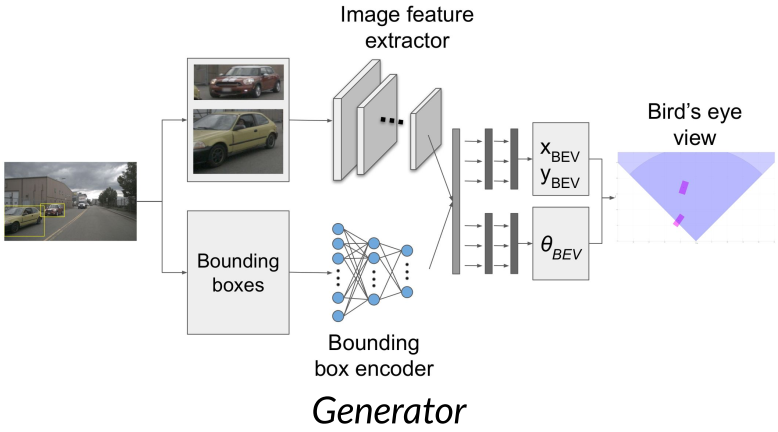
Generator

Input:

- RGB image
- Object bounding boxes

Output:

- BEV representation



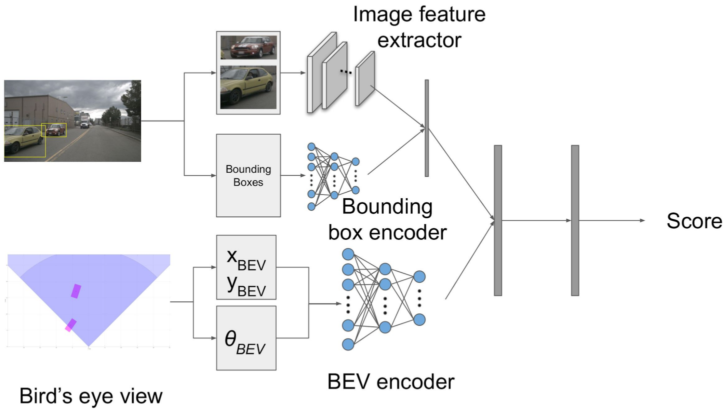
Method
Critic

Input:

- RGB image
- BEV from Generator

Output:

- Score



Critic

Evaluation
Datasets

Trained entirely on KITTI data:

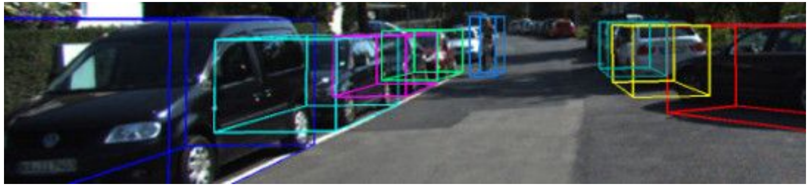
- >13k *'car'* detections

Tested on nuScenes:

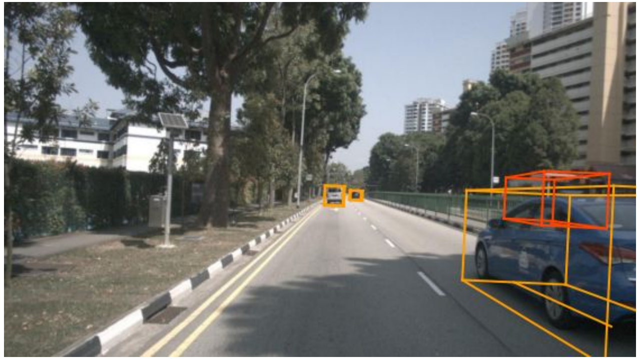
- >60k *'vehicle.car'*

Qualitative evaluation on:

- Virtual KITTI 2
- Surround Vehicle Awareness (GTAV)



KITTI



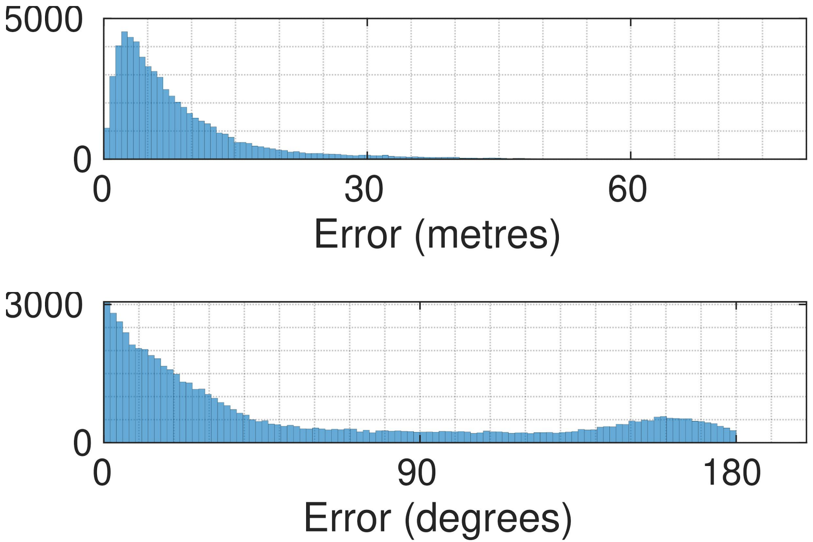
nuScenes

Evaluation

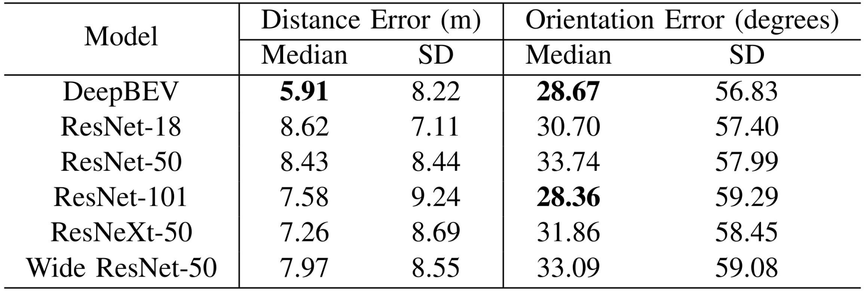
Results

Median Euclidean distance error 22.8% lower than next best

20.6% of the size of ResNet-101



Histogram of per-object errors (nuScenes)

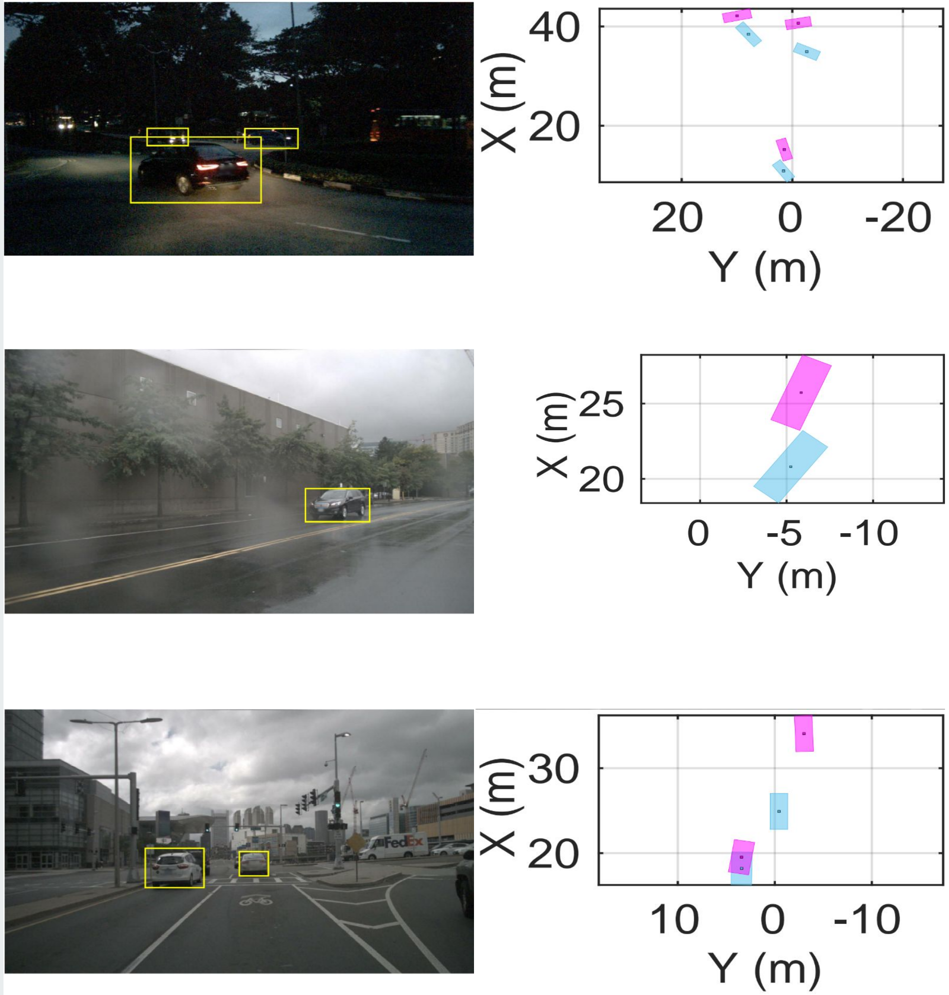


Model	Distance Error (m)		Orientation Error (degrees)	
	Median	SD	Median	SD
DeepBEV	**5.91**	8.22	**28.67**	56.83
ResNet-18	8.62	7.11	30.70	57.40
ResNet-50	8.43	8.44	33.74	57.99
ResNet-101	7.58	9.24	**28.36**	59.29
ResNeXt-50	7.26	8.69	31.86	58.45
Wide ResNet-50	7.97	8.55	33.09	59.08

Results (nuScenes)

Evaluation

Samples



nuScenes

Blue denotes ground truth pose

Magenta denotes model prediction

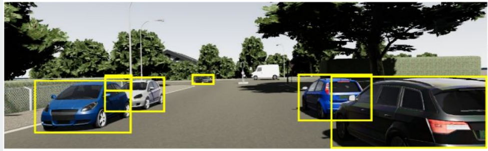
Evaluation

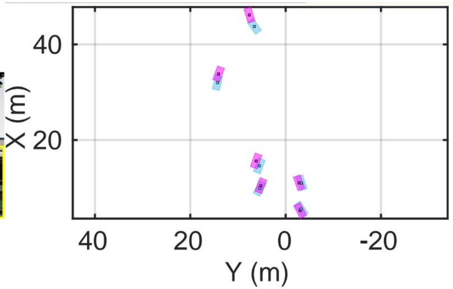
Samples

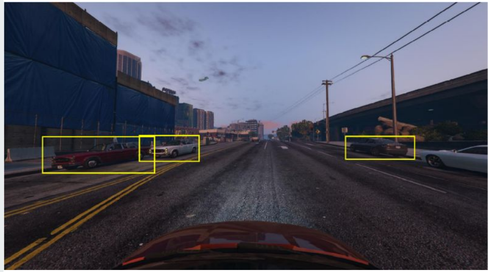
Virtual KITTI 2
Surround Vehicle Awareness (GTAV)

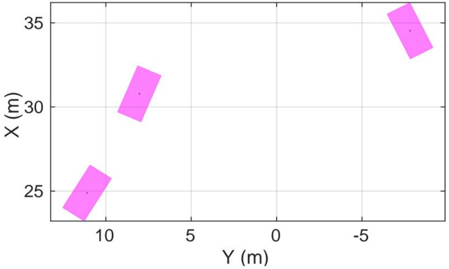
Blue denotes ground truth pose

Magenta denotes model prediction









Conclusion

We demonstrate an adversarial approach to generate BEV representations of a scene from a monocular camera

Adversarial training shows notable improvements on novel data

Thank you for your attention!