PIXOR: Real-time 3D Object Detection from Point Clouds

Bin Yang, Wenjie Luo, Raquel Urtasun
Uber Advanced Technologies Group, University of Toronto

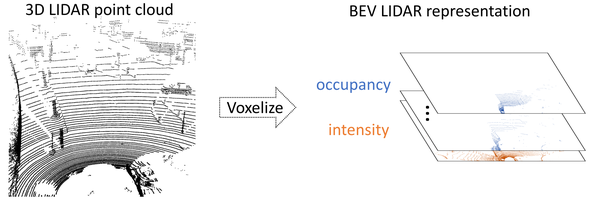
CVPR 2018 SALT LAKE CITY • JUNE 18-22

Summary

- 3D object detection is crucial for autonomous driving.
- LIDAR data is widely used for accurate 3D perception.
- Most LIDAR based 3D detectors run slowly, either because of the 3D LIDAR representation or a two-stage proposal based detection framework.
- **Approach:** ***Single-shot***, ***proposal-free*** detector that operates on bird's eye view (***BEV***) LIDAR representation
- **Performance:** State-of-the-art 3D object detection **(1st on KITTI)** with real-time speed **(~28 FPS)**

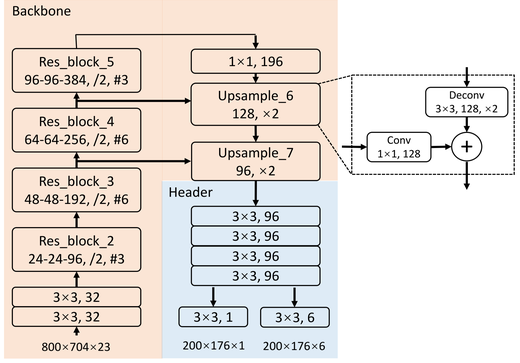
LIDAR Representation

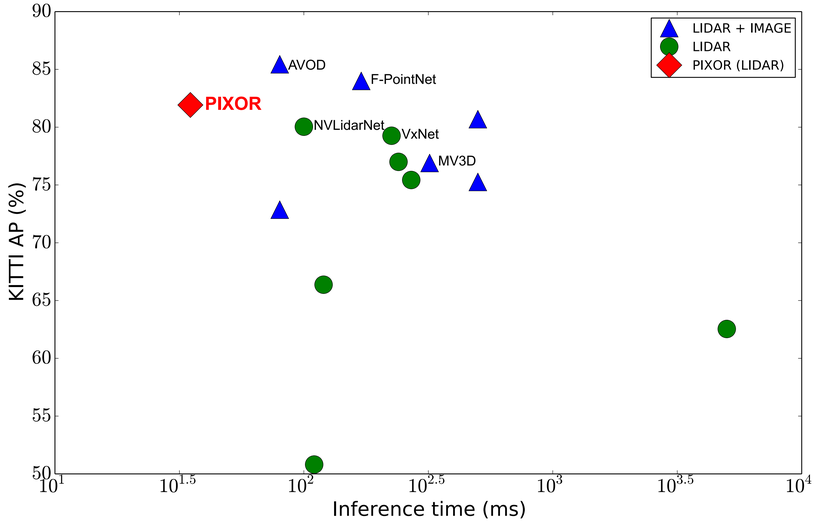
- BEV voxelization: Height as channels



Network Architecture

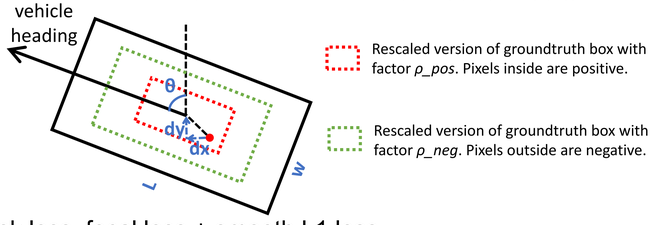
- ResNet backbone with FPN multi-scale feature fusion.
- Fully-convolutional header shared by classification and regression tasks.
- Output pixel-wise dense predictions.
- No pre-trained weights used.





Detection Loss

- Object parameterization: {cos2θ, sin2θ, dx, dy, log(W), log(L)}



- Multi-task loss: focal loss + smooth L1 loss

$$Loss = \text{focal\_loss}(p, y_{cls}) + \text{smooth}_{L_1}(q - y_{reg})$$

$$\text{focal\_loss}(p, y) = \begin{cases} -0.25(1-p)^2 \log(p) & \text{if } y = 1 \\ -0.75p^2 \log(1-p) & \text{otherwise,} \end{cases}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases}$$

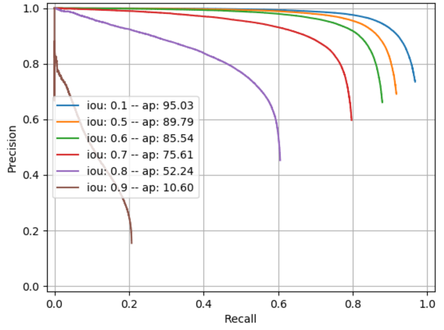
BEV Car Detection on KITTI

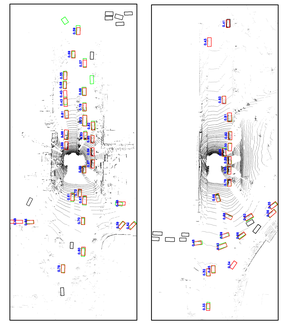
- Dataset: 7,481 frames for training; 7,518 frames for testing.
- Input: X [0, 70m], Y [-40m, 40m], 0.1m resolution
- Runtime ablation on a TITAN Xp GPU:
  - 35 ms = 1ms voxelization + 31ms network + 3ms NMS

Method	Data	Time/ms	AP_mod.	AP_easy	AP_hard
3D FCN	LIDAR	>5000	62.54	69.54	55.94
MV3D	LIDAR	240	77.00	85.82	68.94
VxNet	LIDAR	225	79.26	**89.35**	**77.39**
NVLidarNet	LIDAR	**100**	**80.04**	84.44	74.31
PIXOR	LIDAR	**35**	**81.92**	**87.25**	**76.01**

BEV Car Detection on TOR4D

- TOR4D: a large-scale 3D object detection benchmark collected at Uber ATG with over 1 million frames.
- Training/validation/testing set: 5000/500/1000 video sequences
- Input : X [-100m, 100m], Y [-40m, 40m], 0.2m resolution
- Inference time: 24 ms network on a 1080TI GPU





Conclusion

- 3D detection can be accurate and real-time at the same time!