Multi-View 3D Object Detection Network for Autonomous Driving

Xiaozhi Chen[1] Huimin Ma[1] Ji Wan[2] Bo Li[2] Tian Xia[2]

[1]Tsinghua University [2]Baidu Inc.

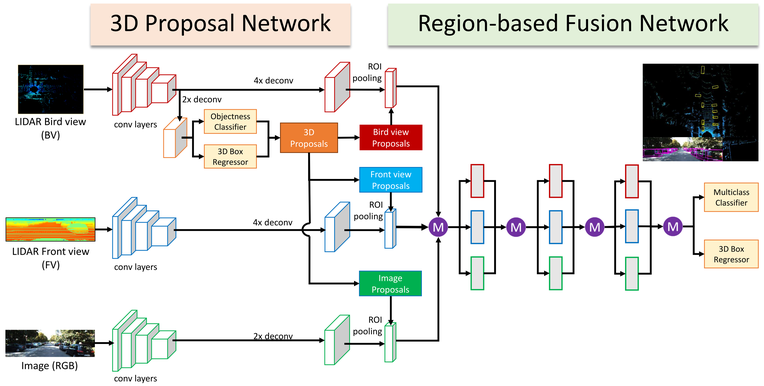
Summary

Goal: 3D object detection vis sensor fusion

Contributions:

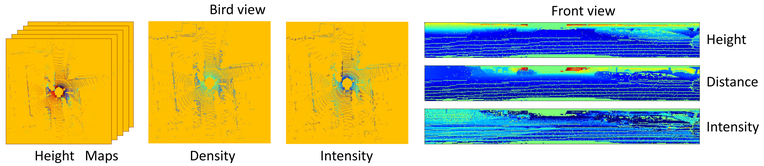
- The first end-to-end 3D detection network for camera and LIDAR fusion
- State-of-the-art 3D box Recall: 99.1% (IoU=0.25), 91% (IoU=0.5)
- Significant improvements in 3D localization (**+25%**), 3D detection(**+30%**) and 2D detection (**+10%**) over previous LIDAR-based methods

MV3D Network



Multi-view Point Cloud Representation:

- sparse 3D point cloud –> compact 2D feature maps
- slow 3D convolutions –> efficient 2D convolutions



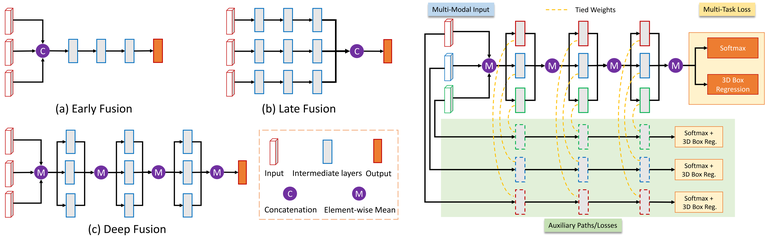
3D Proposal Network:

- 3D box regression based on 2D feature maps
- Multi-View ROI Pooling: establish mappings among multiple views

Region-based Fusion Network

How to enable interactions among multiple views?

- Deep fusion: multi-layer interactions
- Training: drop-path, auxiliary paths/losses



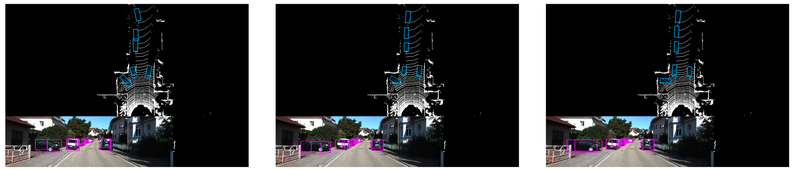
Comparison of Fusion Schemes on KITTI:

Data	AP_{3D} (IoU=0.5)			AP_{loc} (IoU=0.5)			AP_{2D} (IoU=0.7)		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Early Fusion	93.92	87.60	87.23	94.31	88.15	87.61	87.29	85.76	78.77
Late Fusion	93.53	87.70	86.88	93.84	88.12	87.20	87.47	85.36	78.66
Deep Fusion	**96.02**	**89.05**	**88.38**	**96.34**	**89.39**	**88.67**	**95.01**	**87.59**	**79.90**

Ablation Analysis of Features:

Data	AP_{3D} (IoU=0.5)			AP_{loc} (IoU=0.5)			AP_{2D} (IoU=0.7)		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
FV	67.6	56.30	49.98	74.02	62.18	57.61	75.61	61.60	54.29
RGB	73.68	68.86	61.94	77.30	71.68	64.58	83.80	76.45	73.42
BV	92.30	85.50	78.94	92.90	86.98	86.14	85.00	76.21	74.80
FV+RGB	77.41	71.63	64.30	82.57	75.19	66.96	86.34	77.47	74.59
FV+BV	95.19	87.65	80.11	95.74	88.57	88.13	88.41	78.97	78.16
BV+RGB	**96.09**	88.70	80.52	**96.45**	89.19	80.69	89.61	**87.76**	79.76
BV+FV+RGB	96.02	**89.05**	**88.38**	96.34	**89.39**	**88.67**	**95.01**	87.59	**79.90**

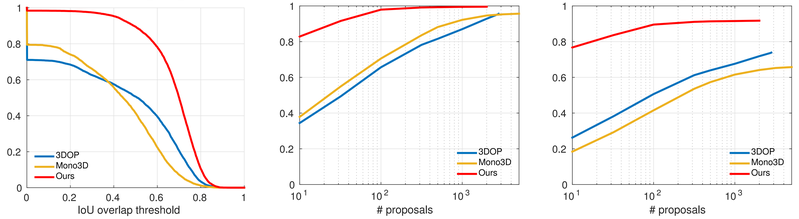
Qualitative Comparisons:



3DOP [NIPS'15] VeloFCN [RSS'16] Ours

KITTI Results

3D Proposal Recall



3D Localization AP on Validation Set (%)

Method	IoU=0.5			IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard
Mono3D [CVPR'16]†	30.50	22.39	19.16	5.22	5.19	4.13
3DOP [NIPS'15]‡	55.04	41.25	34.55	12.63	9.49	7.59
VeloFCN [RSS'16]*	79.68	63.82	62.80	40.14	32.08	30.47
Ours (BV+FV)*	**95.74**	**88.57**	**88.13**	**86.18**	**77.32**	**76.33**
Ours (BV+FV+RGB)*†	**96.34**	**89.39**	**88.67**	**86.55**	**78.10**	**76.67**

3D Detection AP on Validation Set (%)

Method	IoU=0.25			IoU=0.5			IoU=0.7		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Mono3D [CVPR'16]†	62.94	48.2	42.68	25.19	18.2	15.52	2.53	2.31	2.31
3DOP [NIPS'15]‡	85.49	68.82	64.09	46.04	34.63	30.09	6.55	5.07	4.10
VeloFCN [RSS'16]*	89.04	81.06	75.93	67.92	57.57	52.56	15.20	13.66	15.98
Ours (BV+FV)*	**96.03**	**88.85**	**88.39**	**95.19**	**87.65**	**80.11**	**71.19**	**56.60**	**55.30**
Ours (BV+FV+RGB)*†	**96.52**	**89.56**	**88.94**	**96.02**	**89.05**	**88.38**	**71.29**	**62.68**	**56.56**

2D Detection AP on Test Set (%)

Image-based				LIDAR-based			
Method	Easy	Mod.	Hard	Method	Easy	Mod.	Hard
Faster RCNN [NIPS'15]†	87.90	79.11	70.19	Vote3D [RSS'15]*	56.66	48.05	42.64
Mono3D [CVPR'16]†	90.27	87.86	78.09	VeloFCN [RSS'16]*	70.68	53.45	46.90
3DOP [NIPS'15]‡	90.09	88.34	78.79	Vote3Deep [arXiv'16]*	76.95	68.39	63.22
MS-CNN [ECCV'16]†	90.46	88.83	74.76	3D FCN [IROS'17]*	85.54	75.83	68.30
SubCNN [WACV'17]†	90.75	88.86	79.24	Ours (BV+FV)*	**89.80**	**79.76**	**78.61**
SDP+RPN [CVPR'16]†	89.90	89.42	78.54	Ours (BV+FV+RGB)*†	**90.53**	**89.17**	**80.16**

†: Monocular, ‡: Stereo, *: LIDAR