Multi-Task Multi-Sensor Fusion for 3D Object Detection

by

Ming Liang, Bin Yang, Yun Chen, Rui Hu, Raquel Urtasun
Uber ATG & University of Toronto

Presented by
Di Feng

https://frankfengdi.github.io/
Agenda

• Background
  • Sensors for autonomous driving (AD)
  • LiDAR, camera-based scene understanding
  • Multi-task learning (MTL)

• Paper reading (MMF: Multi-task Multi-sensor Fusion)

Background

Sensor suite for autonomous vehicles


Source: https://velodynelidar.com/, https://www.bosch.com/
Background
LiDAR and camera, a comparison

<table>
<thead>
<tr>
<th></th>
<th>LiDAR</th>
<th>RGB camera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Texture</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Lightning</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Weather</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Research question: How to combine LiDAR and camera data?

Source: https://waymo.com/open/data/
Background
Multi-task learning

Semantic segmentation
Depth estimation
Road estimation
Object detection

And many more, e.g. tracking, scene flow, traffic light detection, prediction ...

Research question: How to combine multiple perception tasks together?

Source: [http://www.cvlibs.net/datasets/kitti/](http://www.cvlibs.net/datasets/kitti/)
Background
Multi-task learning using deep neural networks

- One model, multiple tasks ("swiss army knife")
- Benefits: higher efficiency, lower memory footprint, better accuracy
- Challenge: negative transfer, # tasks, loss weight, catastrophic forgetting

MMF (Multi-task Multi-sensor Fusion)

Overview

- Fuse LiDAR + Camera data via a CNN
- Leverage ground estimation and depth completion (auxiliary tasks) for 2D/3D detection (target tasks)
- Full model is end-2-end and trainable
MMF

Overview
MMF

Fusion Network Architecture

LiDAR Point Cloud → Ground-Relative BEV Representation → LiDAR Backbone Network

Residual block + FPN

LiDAR Bird’s Eye View

2 fcl (256 units)

Feature concat.

ROI refinement in typical two-stage detection pipeline

Sparse Depth Image → Concat → RGB Image

Dense Fusion
MMF

Fusion Network Architecture: Dense fusion

MMF

Auxiliary task: ground estimation (GE)

- To remove ground points, and normalize bounding boxes’ height by subtraction
- A regression task, estimate ground height at each voxel in the BEV space
- Small U-Net architecture (8ms / frame)

Comments

- GE could also help SLAM for ground removal
- Small network is enough
- GE only as input -> why not in output, like standard MTL?


Bin Yang, Ming Liang, and Raquel Urtasun. HDNet: Exploiting hd maps for 3d object detection. In *CoRL*, 2018
Auxiliary task: depth completion (DC)

- Predict depth for each image pixel -> create pseudo-LiDAR points
- Densify LiDAR points with sparse LiDAR points + pseudo-LiDAR points
- Help detection especially at long range

Wang, Y., Chao, W. L., Garg, D., Hariharan, B., Campbell, M., & Weinberger, K. Q.
Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In CVPR, 2019
MMF

Implementation details

• End-to-end training
• Multi-task learning with fixed loss weights
• Pre-trained U-Net for ground estimation, then fine-tuning implicitly
• Train and evaluate on TOR4D (unpublished) and open KITTI dataset
• 80 ms / frame (pretty good!)

\[ \text{Loss} = L_{cls} + \lambda (L_{box} + L_{r2d} + L_{r3d}) + \gamma L_{depth} \]

Comments

• Tricky for ground estimation -> implicit training
• Hyper-parameters is difficult to tune in such a large network
Experimental results (on KITTI)

- LiDAR-camera fusion helps a lot
- Second-stage bbox refinement is important
- Online mapping contributes to <2% AP gains
- Depth completion and fusion contributes to <1% AP gains

<table>
<thead>
<tr>
<th>Model</th>
<th>Multi-Sensor</th>
<th>Multi-Task</th>
<th>2D AP (%)</th>
<th>3D AP (%)</th>
<th>BEV AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pt roi map</td>
<td>dep depf</td>
<td>easy mod.</td>
<td>hard</td>
<td>easy mod.</td>
</tr>
<tr>
<td>LiDAR only</td>
<td></td>
<td></td>
<td>93.44</td>
<td>87.55</td>
<td>84.32</td>
</tr>
<tr>
<td>+image</td>
<td>✓</td>
<td>✓</td>
<td>+2.95</td>
<td>+1.97</td>
<td>+2.76</td>
</tr>
<tr>
<td>+map</td>
<td>✓</td>
<td>✓</td>
<td>+3.06</td>
<td>+2.20</td>
<td>+3.33</td>
</tr>
<tr>
<td>+refine</td>
<td>✓</td>
<td>✓</td>
<td>+3.94</td>
<td>+2.71</td>
<td>+4.66</td>
</tr>
<tr>
<td>+depth</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+4.69</td>
</tr>
<tr>
<td>full model</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>


MMF (Multi-task Multi-sensor Fusion)

Discussion

• A LiDAR-camera fusion network for
  • Depth completion: densify LiDAR points
  • Ground estimation: remove LiDAR points on ground and normalize detection
  • 2D/3D detection

• **Pros**: shows how to leverage multi-modal sensor and auxiliary tasks to improve detection

• **Cons**: Too complex network architecture design and training -> a unified MTL is expected

• Future directions:
  • More compact and simple MTL
  • More auxiliary tasks (i.e. road segmentation, scene flow, lanes, lidar semseg … )