Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation

Uber ATG

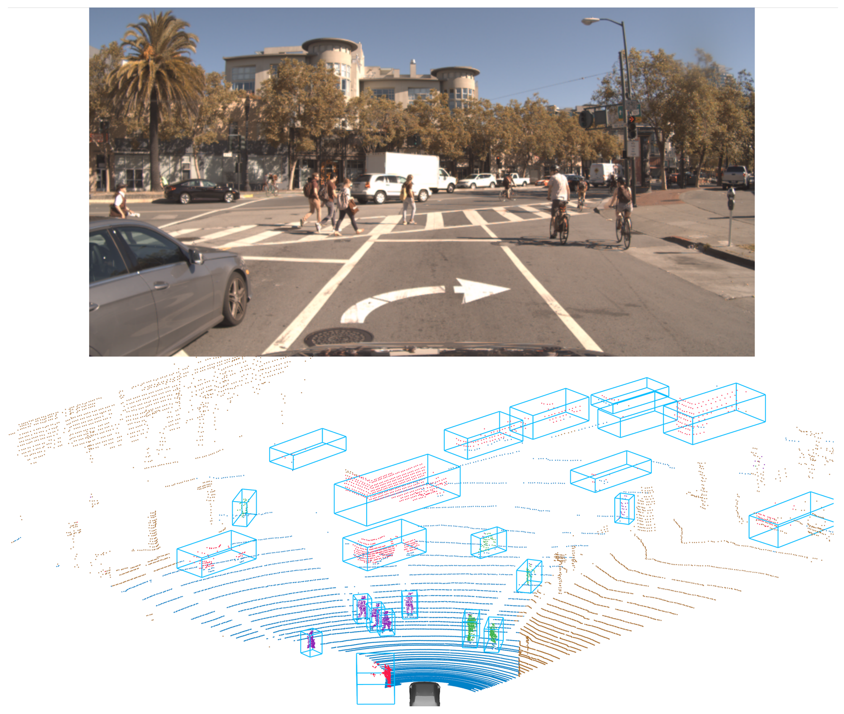
Gregory P. Meyer gmeyer@uber.com
Jake Charland jakec@uber.com
Darshan Hegde darshan.hegde@uber.com
Ankit Laddha aladdha@uber.com
Carlos Vallespi-Gonzalez cvallespi@uber.com

Uber Advanced Technologies Group

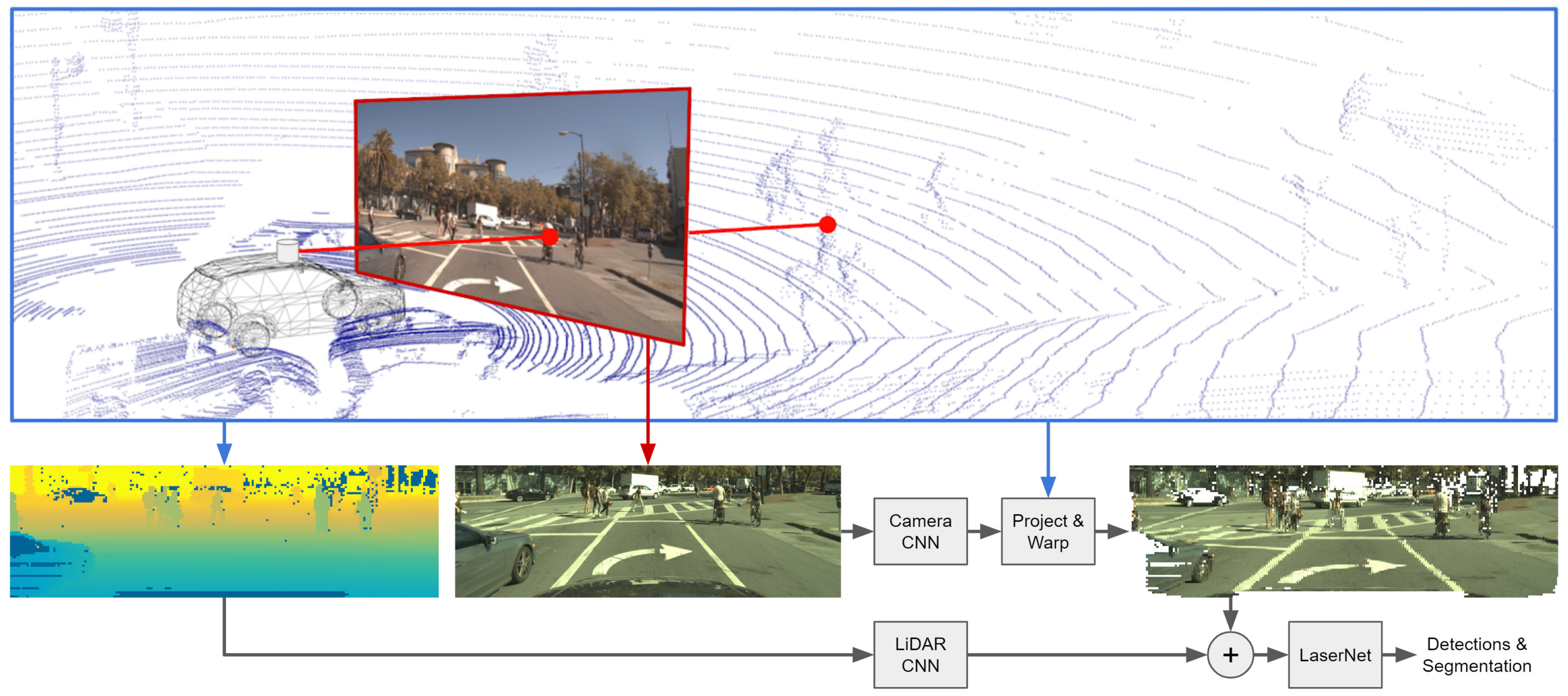
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

- 3D object detection and semantic scene understanding are two fundamental capabilities for autonomous driving.
- We present a method for fusing 2D image data and 3D LiDAR data, and we leverage this approach to improve LaserNet, a LiDAR based 3D object detector.
- Additionally, we extend the model to perform 3D semantic segmentation.
- On a large dataset, our approach achieves state-of-the-art performance on both object detection and semantic segmentation.
- Our extensions are lightweight, adding only 8 ms to the runtime of LaserNet.



Overview



- Our method fuses 2D camera images and 3D LiDAR measurements.
- Both sensor modalities are represented as images, where the 3D data is represented using the native range view of the LiDAR.
- Our approach associates LiDAR points with camera pixels by projecting the 3D points onto the 2D image, and this mapping is used to warp information from the camera image to the LiDAR image.
- Instead of warping RGB values as shown, we fuse features extracted by a CNN.
- The LiDAR and camera features are concatenated and passed to LaserNet.
- The entire model is trained end-to-end to perform 3D object detection and semantic segmentation without the need for additional image labels.

LaserNet

- *LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving* will be present at the main conference during the Thursday afternoon poster session.
- LaserNet was also developed by members of the Uber Advanced Technologies Group.
- For more information on LaserNet, please visit our poster, **#209**.

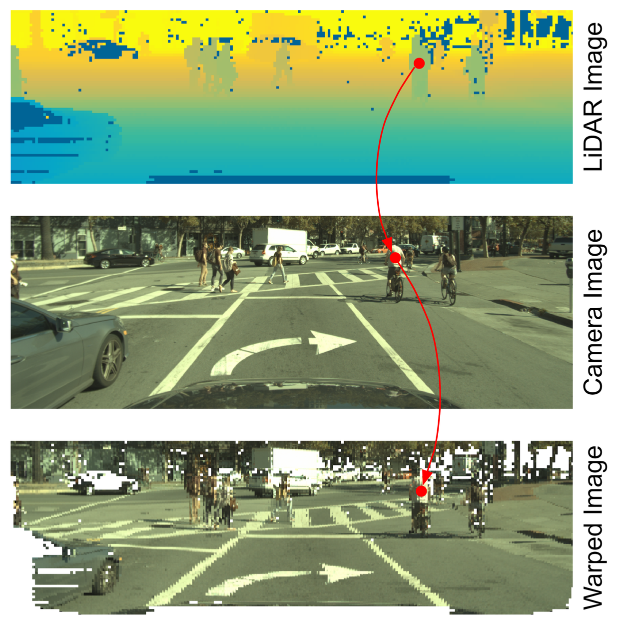
Method

1 Sensor Fusion

- The 2D image and the 3D points are related through projective geometry.
- To fuse the LiDAR and RGB data, each LiDAR point $\boldsymbol{p}$ is projected onto the RGB image:

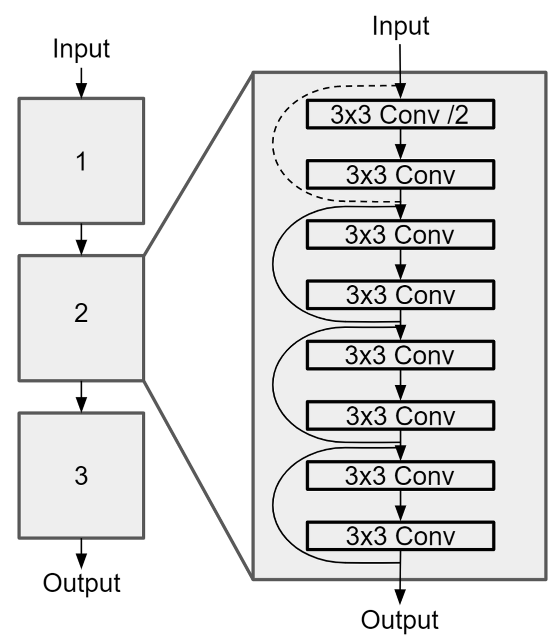
$$\alpha\,[u, v, 1]^T = \boldsymbol{K}\,(\boldsymbol{R}\boldsymbol{p} + \boldsymbol{t})$$

- This provides a mapping from the LiDAR image to the RGB image, which is used to copy features from the RGB image into the LiDAR image.



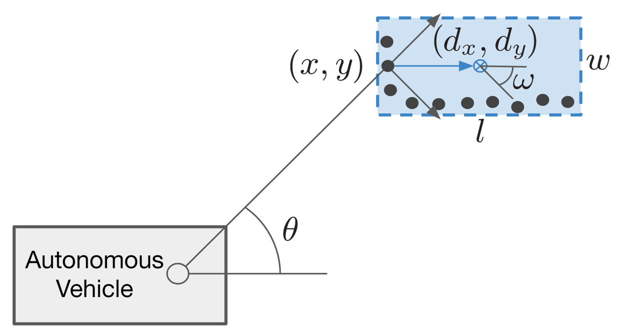
2 Network Architecture

- Fusing raw RGB data would result in a significant amount of information loss. Instead, we fuse features extracted by a CNN from the RGB image.
- The image network contains three ResNet blocks, where each block downsamples the feature map by half and performs a set of 2D convolutions.
- The image features are warped into the LiDAR image and concatenated with the LiDAR features then passed to LaserNet.
- The image network is trained by back-propagating the loss through the warped image features.



3 Predictions

- The network is trained to predict a set of class probabilities for each point in the image.
- Given a point is on an object, the network estimates the bounding box by predicting a center, orientation, and dimensions relative to the point.



Detection Results

- Our approach is evaluated and compared to state-of-the-art methods in both 3D object detection and semantic segmentation on the large-scale ATG4D dataset.
- The dataset contains 5,000 sequences for training and 500 sequences for validation.
- The detection and segmentation performance of our method and the existing work is evaluated within the front 90° field of view and up to 70 meters away.

Table 1: BEV Object Detection Performance

Method	Input	Vehicle $AP_{0.7}$	Bike $AP_{0.5}$	Pedestrian $AP_{0.5}$
PIXOR	LiDAR	80.99	-	-
PIXOR++	LiDAR	82.63	-	-
ContFuse	LiDAR	83.13	57.27	73.51
LaserNet	LiDAR	85.34	61.93	80.37
ContFuse	LiDAR+RGB	85.17	61.13	76.84
LaserNet++ (Ours)	LiDAR+RGB	**86.23**	**65.68**	**83.42**

- On the ATG4D dataset, our approach achieves state-of-the-art performance.
- Adding the supplemental 2D data improves performance on smaller objects (pedestrian and bike) where the LiDAR receives fewer measurements.

Segmentation Results

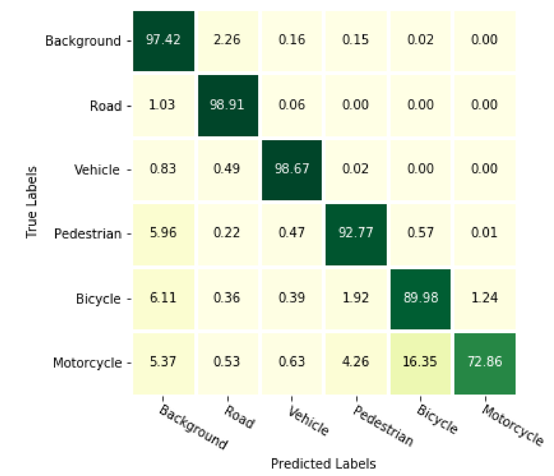
Table 2: 3D Semantic Segmentation Performance

Method	Input	mAcc	mIoU
2D U-Net	LiDAR	81.95	76.39
LaserNet++ (Ours)	LiDAR+RGB	**91.77**	**86.62**

Table 3: Per-Class Semantic Segmentation Performance

Method	Class IoU					
	Background	Road	Vehicle	Pedestrian	Bicycle	Motorcycle
2D U-Net	92.03	97.92	93.76	74.47	61.25	38.90
LaserNet++ (Ours)	**93.59**	**98.23**	**97.67**	**86.19**	**80.98**	**63.07**

- To perform semantic segmentation, we classify each point in the LiDAR image with its most likely class according to the predicted class probabilities.
- On this dataset, our approach considerably outperforms the existing method across all metrics.
- LaserNet++ performs particularly well on smaller classes (pedestrian, bicycle, and motorcycle).



Examples