HD-Map Fusion with Deep Learning for Object Detection

Onur Can Yücedağ onur@adastec.com

Emre Erdem eerdem95@hotmail.com

Dr. Mehmet Kadir BARAN

Introduction

The Problem:

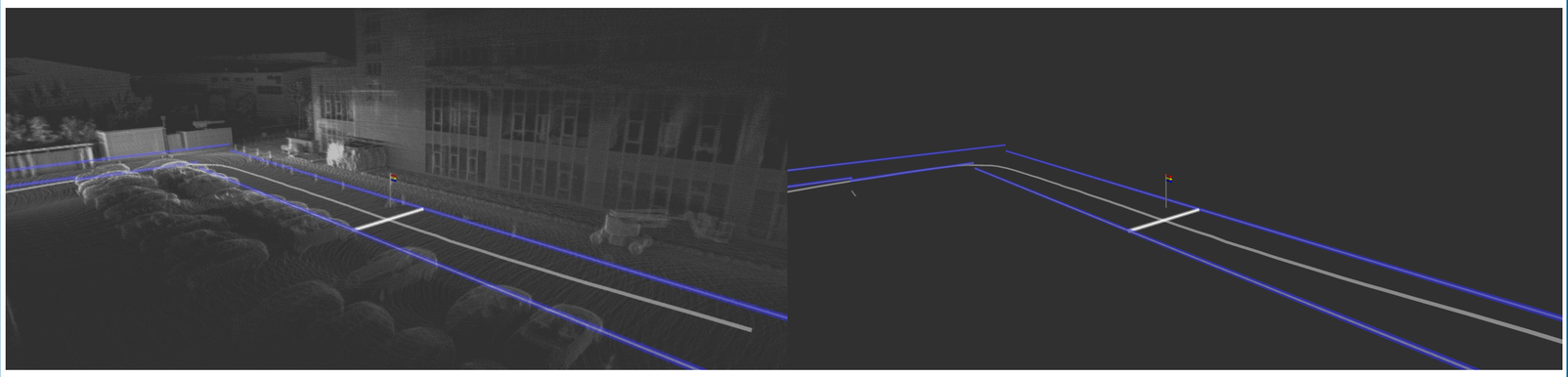
A dramatic increase in computation power allowed us to compute complex object detection problems with advanced Convolutional neural network techniques. One of the hottest sectors for object detection is their use for autonomous vehicles. This is also one of the hardest sectors because the model is expected to work at least in real-time and work reliably enough to let this information affect the trajectory estimations of an ego vehicle.

Our Approach To The Problem:

We modify already existing object detection models with sensor fusion techniques using vector maps to achieve better results in Region proposal models such as R-CNN family and famous one stage models such as YOLO.

What are Vector Maps?

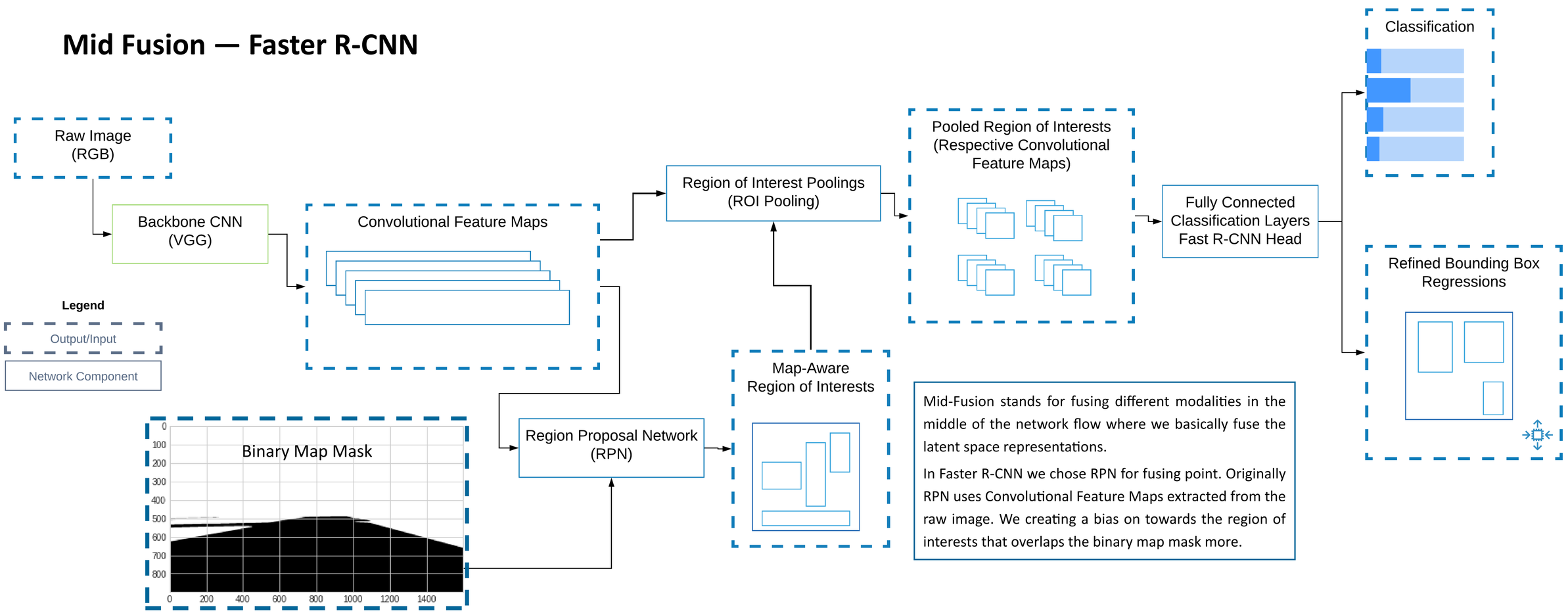
High Definition Map (HD-Map) data that have high precision at centimeter-level. This type of information is mostly used in autonomous cars to understand the important objects in the maps. Differentiating between traffic objects, rules and most importantly the road, sidewalk information is a key component in our sensor fusion implementation. We are using these 3D lane polygons by projecting them on the camera frame.



Vector Map (Traffic light, stop line, curbs, lanes) with point cloud data

Vector Map (Traffic light, stop line, curbs, lanes) data on its own

Mid Fusion — Faster R-CNN



Mid-Fusion stands for fusing different modalities in the middle of the network flow where we basically fuse the latent space representations.

In Faster R-CNN we chose RPN for fusing point. Originally RPN uses Convolutional Feature Maps extracted from the raw image. We creating a bias on towards the region of interests that overlaps the binary map mask more.

Faster R-CNN Test Results

Faster R-CNN | nuScenes Dataset

RoI	Overlap Coefficient	Secs/Image	Car mAP
Top 300	0*	0.420	0.2875
	0**	1.329	0.2868
	1	1.290	0.2859
	2	1.273	0.2869
	3	1.302	0.2870
Top 50	0*	0.395	0.2853
	0**	1.271	0.2855
	1	1.191	0.2775
	2	1.193	0.2509
	3	1.190	0.2505
Top 20	0*	0.321	0.2492
	0**	1.174	0.2492
	1	1.177	0.2424
	2	1.184	0.2375
	3	1.160	0.2221

*: Don't calculate map overlap

**: Calculate map overlap but don't use it

YOLOv3

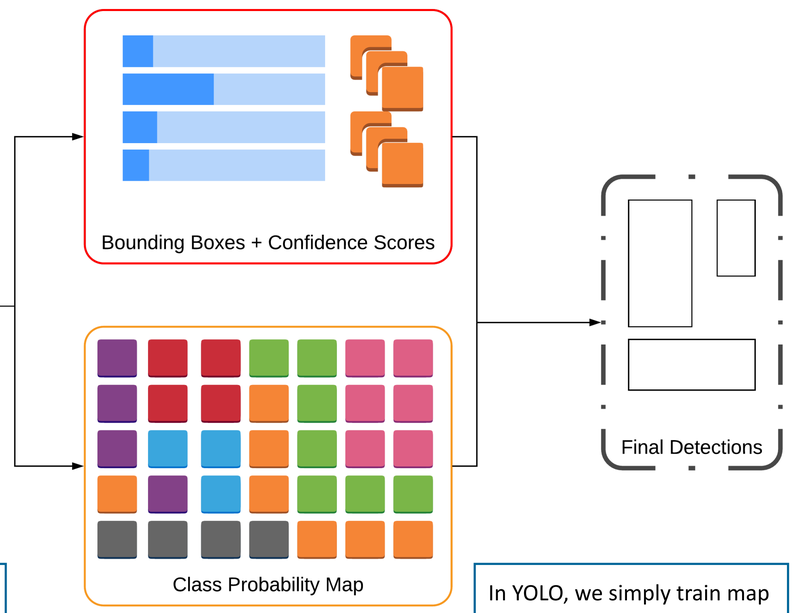
Training Dataset	Validation Dataset	Pretrained Dataset	Car mAP
Original nuScenes	nuScenes Mini	COCO+ImageNet	66.41
Map Overlay nuScenes	nuScenes Mini	COCO+ImageNet	65.99

Faster R-CNN | VOC2007 Dataset

RoI	Secs/Image	Car mAP
Top 300	0.420	0.2875
Top 50	1.329	0.2868

Early Fusion — YOLO

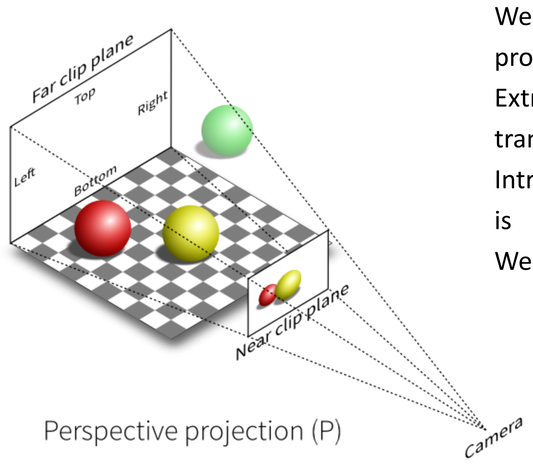
Map Overlay Image



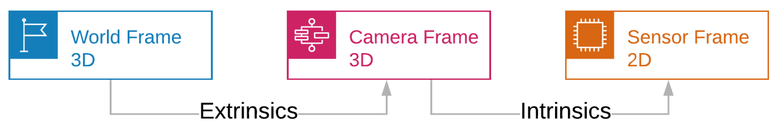
Early-Fusion stands for fusing different modalities in the beginning of the network flow, so we are fusing the raw data from both modalities.

In YOLO, we simply train map overlay RGB images.

Perspective Projection



We used camera calibration pipeline for creating an map overlay image. Camera calibration process generates respective Intrinsic Matrix and Extrinsic Matrix. Extrinsic Matrix is rigid body transformation in homogeneous coordinates which used for transforming an object location in world frame (3D) to camera frame(3D). Intrinsic Matrix contains focal length, principal points and skew coefficient for the camera. It is used for projecting an object in camera frame (3D) to sensor frame (2D). We are getting the vector map as polygons. Once we have the polygon we are projecting the



References

- Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).
- Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
- H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving", In arXiv preprint arXiv:1903.11027.

Onur Can Yücedağ

onur@adastec.com