UNIVERSITY OF Southampton
School of Electronics and Computer Science

Fusion
(I)

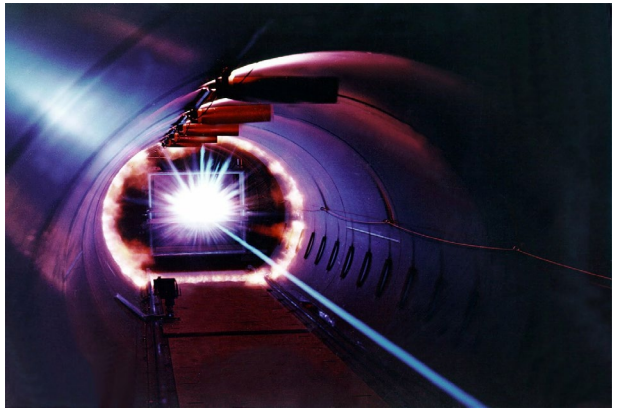
Runwei Guan (MSc Data Science)
thinkerai@foxmail.com / rg6n20@soton.ac.uk
03 / August / 2021

Where are the fusion data from?

Multiple Sensors

Millimeter-wave Radar

- High Stability
- Long Measurement Distance
- High measurement Accuracy
- Low Resolution
- Unable to identify target feature information



LIDAR

- 3D Information Detection
- High Detection Accuracy
- High Resolution
- Poor Stability
- High Cost
- Short Life Span

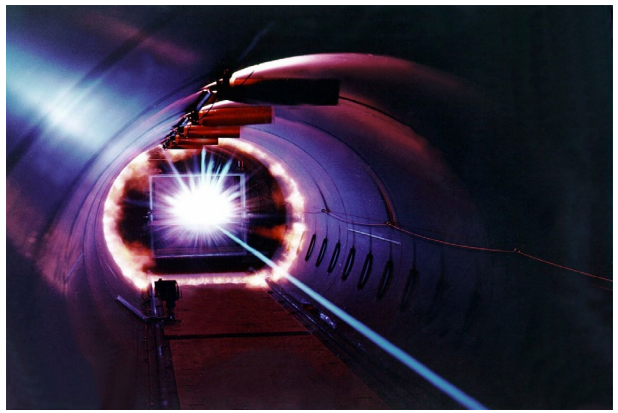
Single Vision Camera Sensor

- 3D Information Detection
- Relative High Detection Accuracy
- Relative High Resolution
- Poor Stability
- Short Longitudinal Detection Distance Range

Why fusion?

Ensemble Weak Models

+



OR

- 3D Information Detection
- High Detection Accuracy
- Long Measurement Distance
- High Resolution
- Relative Poor Stability
- Relative High Cost
- Relative Short Life Span

Why fusion?

Ensemble Weak Models

+

OR

- 3D Information Detection
- High Detection Accuracy
- Long Measurement Distance
- High Stability
- Low Cost
- Long Life Span
- Relative Low Resolution

How fusion?

Fusion Patterns

- Fusion of Raw Data
- Fusion of Proposal
- Fusion of Feature
- Fusion of Target

How fusion?

Fusion of Raw Data

Key

Let the radar point cloud data coordinate target be projected on the image pixel, and the image pixel is jointly calibrated and matched.

Shortcoming

The radar resolution is low, the number of point clouds is very small, and the noise is large, it is difficult to match the image.

How fusion?

Fusion of Proposal

Key

Project the radar point target onto the image, around the point we generate a matrix of interest area, and then we only search in this area, and after the search is found to match the radar point target.

- Quickly eliminate a large number of areas that will not have the target, and greatly improve the recognition speed.
- Non-target targets detected by radar can be quickly eliminated, enhancing the reliability of the results.

Shortcoming

Ideally, the radar point appears in the middle of the vehicle. First of all, because the lateral distance of the target provided by the radar is not accurate, coupled with the error of the camera calibration, the deviation of the projection point of the radar to the target may be serious.

How fusion?

Fusion of Feature

Key

It uses different independent networks for different modalities to extract feature maps. Then, the bounding boxes are proposed separately for each modality and are combined later.

Having an image-based object proposal network in addition to the radar-based network improves the object detection accuracy, as they complement each other by using two different modalities for proposal generation and distance estimation.

Shortcoming

Which to believe?

How fusion?

Fusion of Target

Key

Effectively merge the obstacle results detected by the image with the results detected by the radar.

Shortcoming

The longitudinal distance recognized by the monocular camera is not accurate, and it is difficult to match accurately when there are many obstacles. It may also be recognized by the radar, but not by the monocular camera.

Related Literature

1) Chen X, Ma H, Wan J, et al. Multi-view 3d object detection network for autonomous driving[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 1907-1915.

2) Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 1-8.

3) Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 918-927.

4) Nabati R, Qi H. Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles[J]. arXiv preprint arXiv:2009.08428, 2020.

5) Lim T Y, Ansari A, Major B, et al. Radar and camera early fusion for vehicle detection in advanced driver assistance systems[C]//Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems. 2019.

Related Work and Improvement

3D Object Detection in Point Cloud

Method of current mainstream:

- Encode 3D point cloud with grid representation. Sliding Shapes and Vote3D apply SVM classifiers on 3D grids encoded with geometry features.
- Some recently proposed methods improve feature representation with 3D convolution networks, which, however require expensive computations.

Shortcoming:

Expensive computations

Method of this paper:

- Encode 3D point cloud with multi-view feature maps, enabling region-based representation for multimodal fusion.

Related Work and Improvement

3D Object Detection in Images

Method of current mainstream:

- Use a series of ACF detectors to do 2D detection and 3D pose estimation through 3D voxel mode (such as 3DVP) .

Shortcoming:

Image-based methods usually rely on accurate depth estimation or marker detection.

Method of this paper:

- Fusion of radar point clouds to improve the effect of 3D localization.

Related Work and Improvement

Multimodal Fusion

Method of current mainstream:

- Combine images, depth and optical flow using a mixture-of-experts framework for 2D pedestrian detection.

Shortcoming:

There is too little work in this area, and the method development is not perfect.

Method of this paper:

- Design a deep fusion approach inspired by FractalNet and Deeply-Fused Net.
- In FractalNet, a base module is iteratively repeated to construct a network with exponentially increasing paths.
- Deeply-Fused Net is strutted by combining shallow and deep subnetworks.
- Network in this paper differs from them by using the same base network for each column and adding auxiliary paths and losses for regularization.

Related Work and Improvement

3D Object Proposals

Method of current mainstream:

- Design some depth features based on the stereo point cloud to generate some 3D candidate frames (such as 3DOP), or use the ground plane and some semantic information to generate 3D candidate regions (such as Mono3D).

Shortcoming:

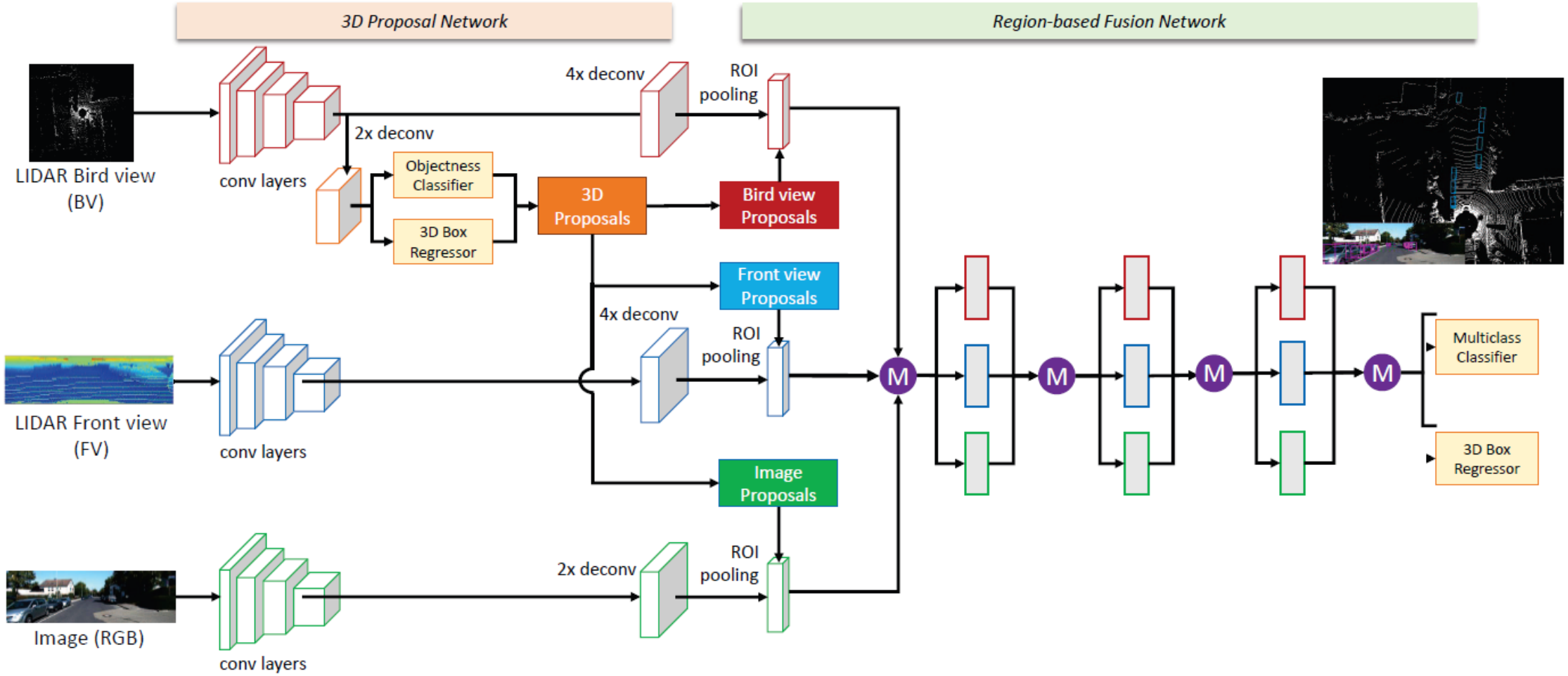
Both 3DOP and Mono3D use manual features.

Method of this paper:

- Use the top view representation of the point cloud and apply 2D convolution to generate 3D candidate regions.

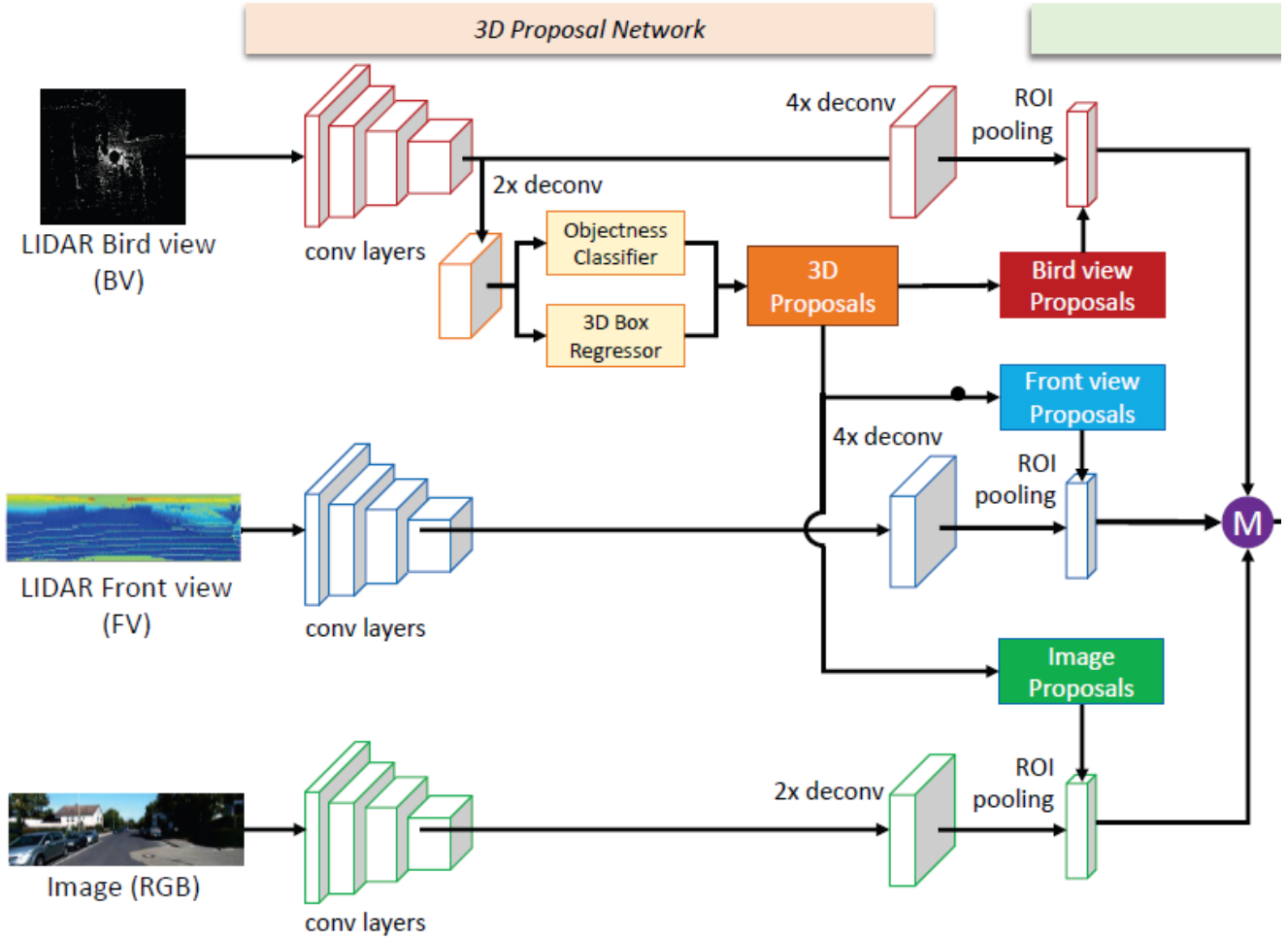
MV3D

Network Structure



MV3D

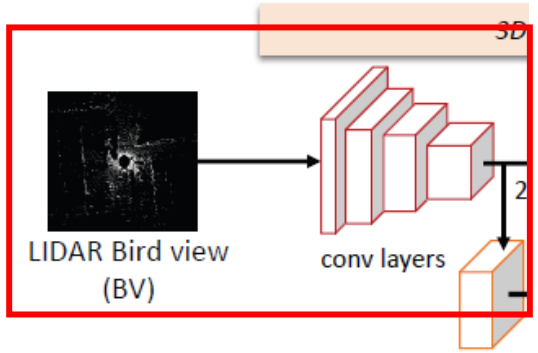
Main Part of Network

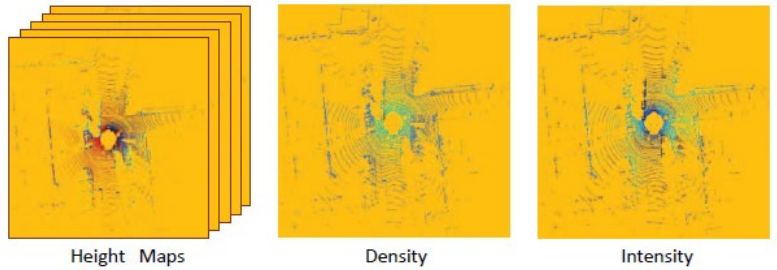


MV3D

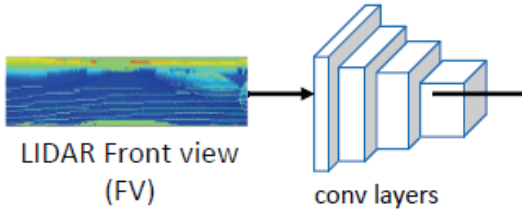
Feature Extraction

Bird View of Point Cloud

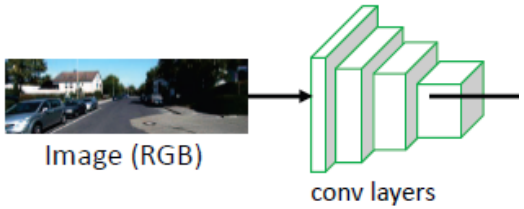




Front View of Point Cloud



Front View of Camera



Height Record the maximum height of all points in each grid as the height feature. In order to encode more height features, the point cloud is divided into M blocks, and the corresponding height map is calculated for each block, thereby obtaining M height maps.

Density $min(1.0, \frac{log(N+1)}{log(64)})$

Intensity The mapped value of the point with the largest height in each cell.

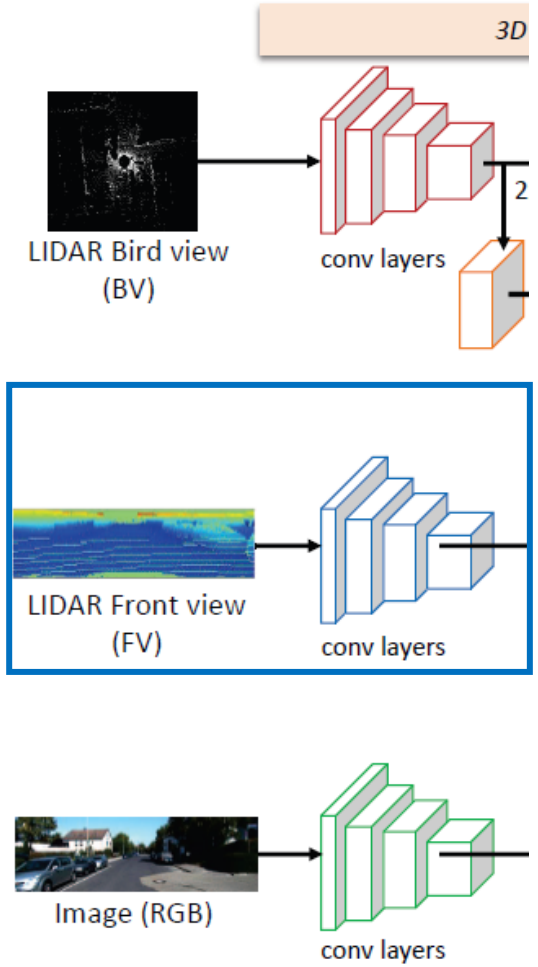
MV3D

Feature Extraction

Bird View of Point Cloud

Front View of Point Cloud

Front View of Camera

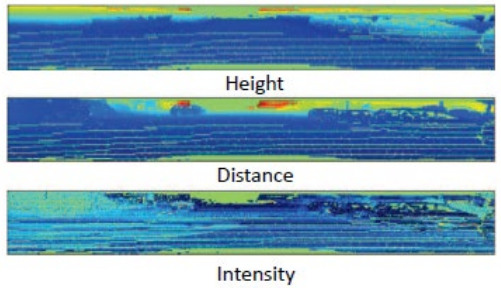


$$p = (x, y, z)$$

$$p_{fv} = (r, c)$$

$$c = \lfloor \text{atan2}(y, x)/\Delta\theta \rfloor$$

$$r = \lfloor \text{atan2}(z, \sqrt{x^2 + y^2})/\Delta\phi \rfloor$$



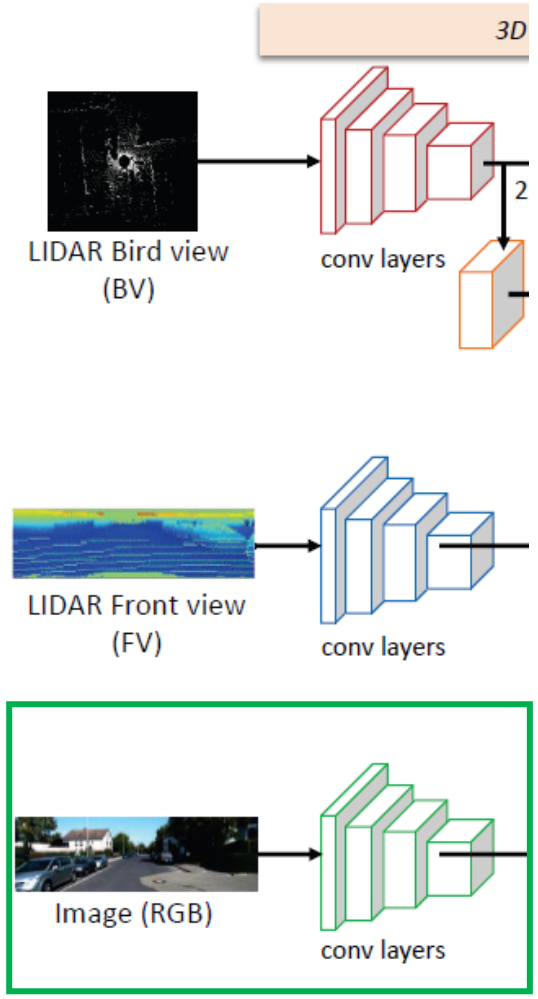
MV3D

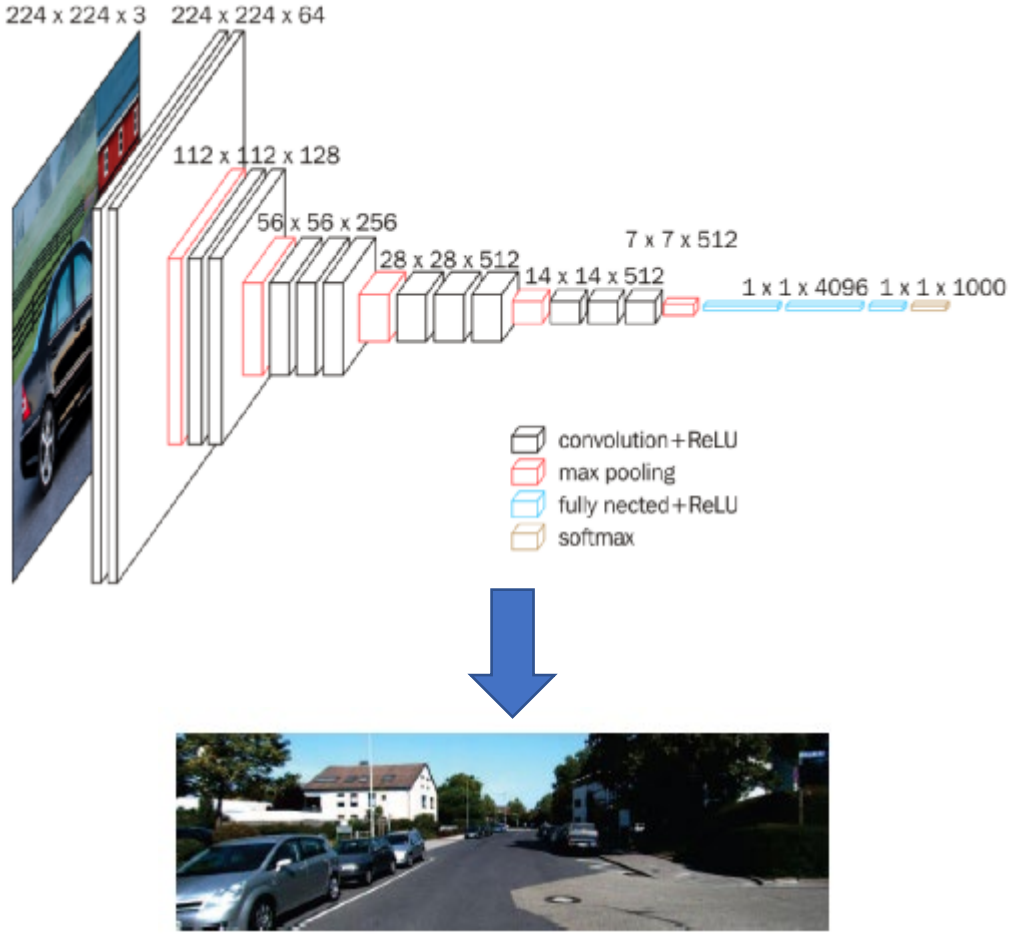
Feature Extraction

Bird View of Point Cloud

Front View of Point Cloud

Front View of Camera



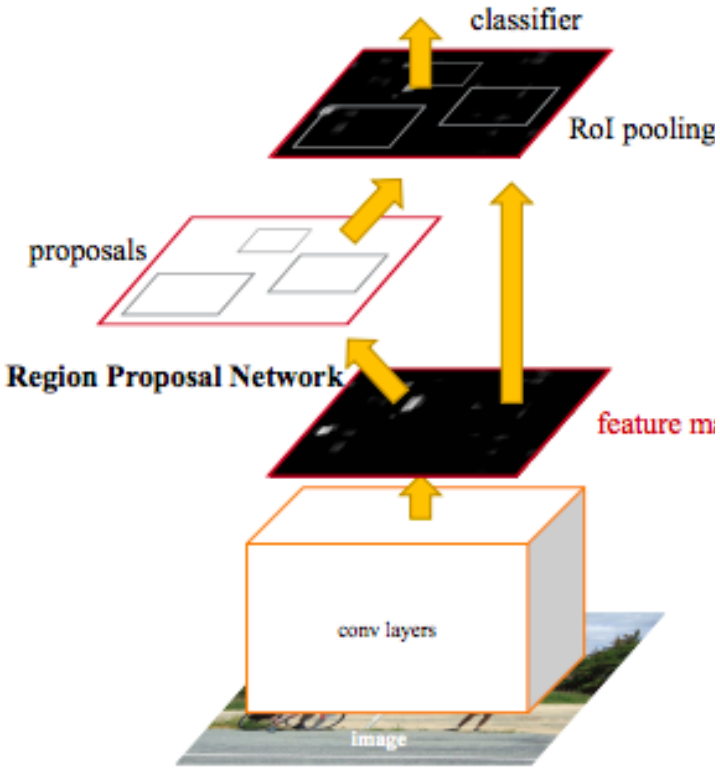


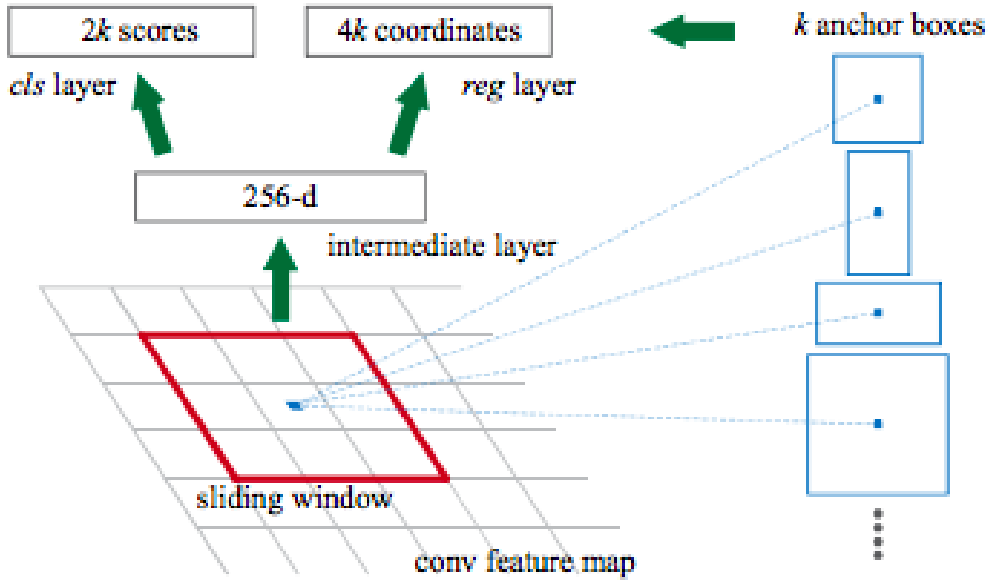
MV3D

3D Proposal Network

Input: Feature Map

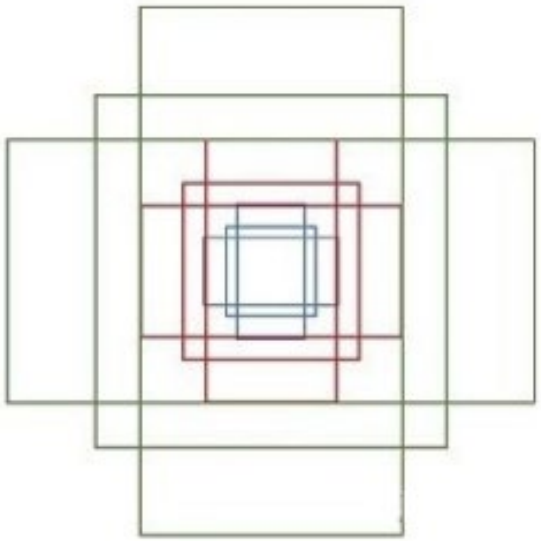
Output: Proposal

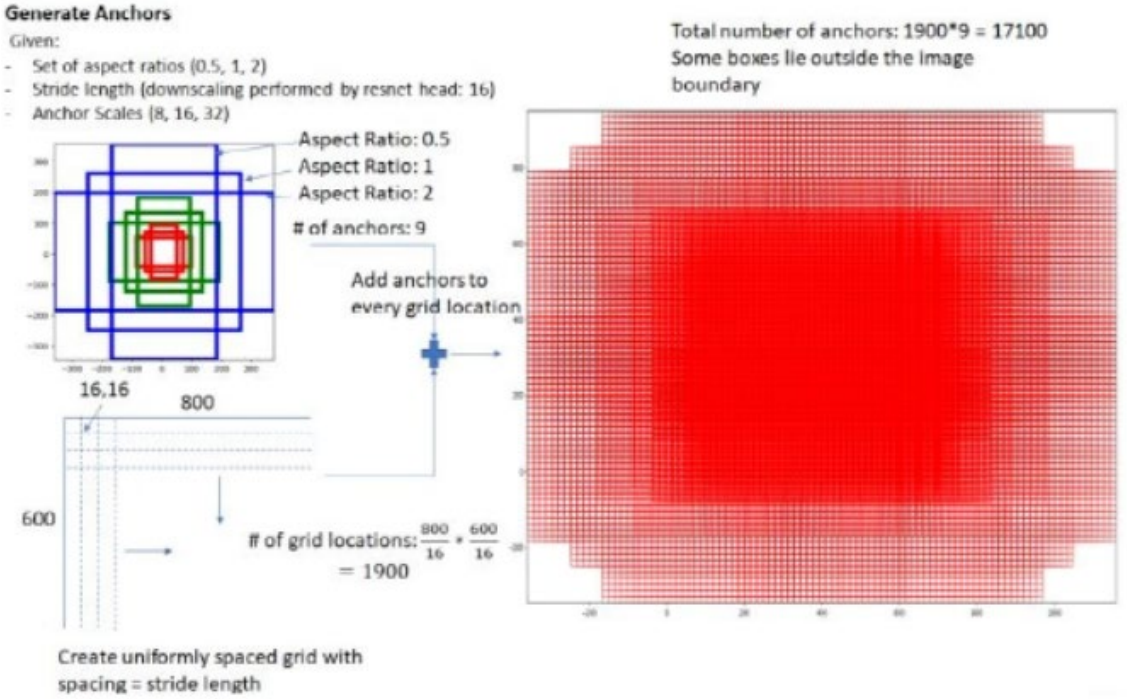


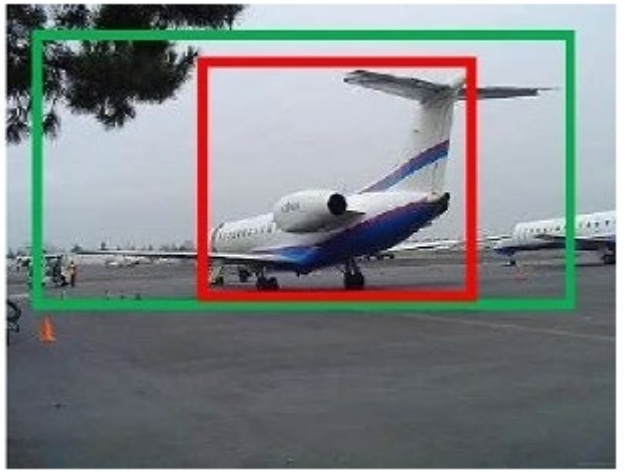


MV3D

3D Proposal Network

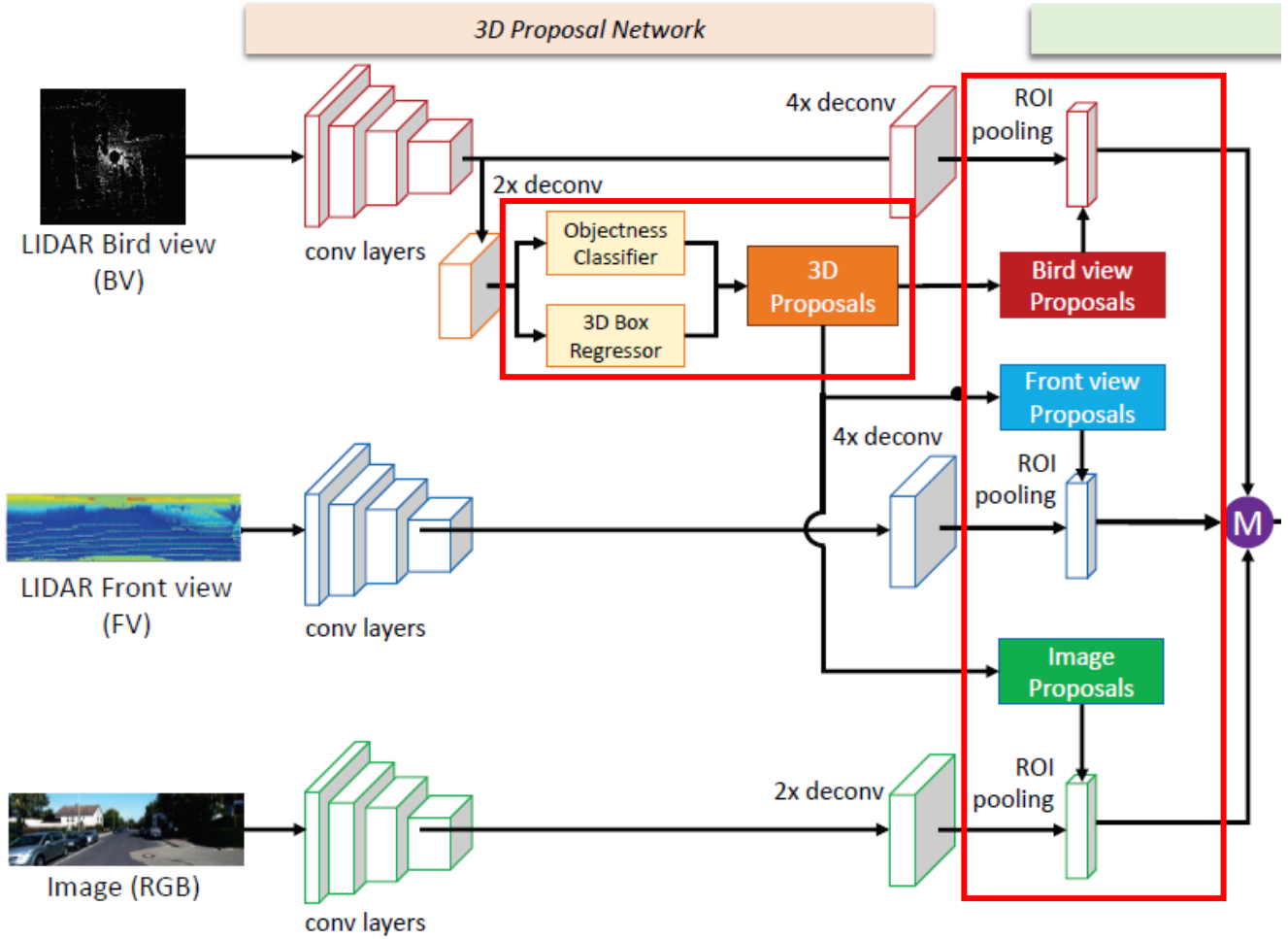






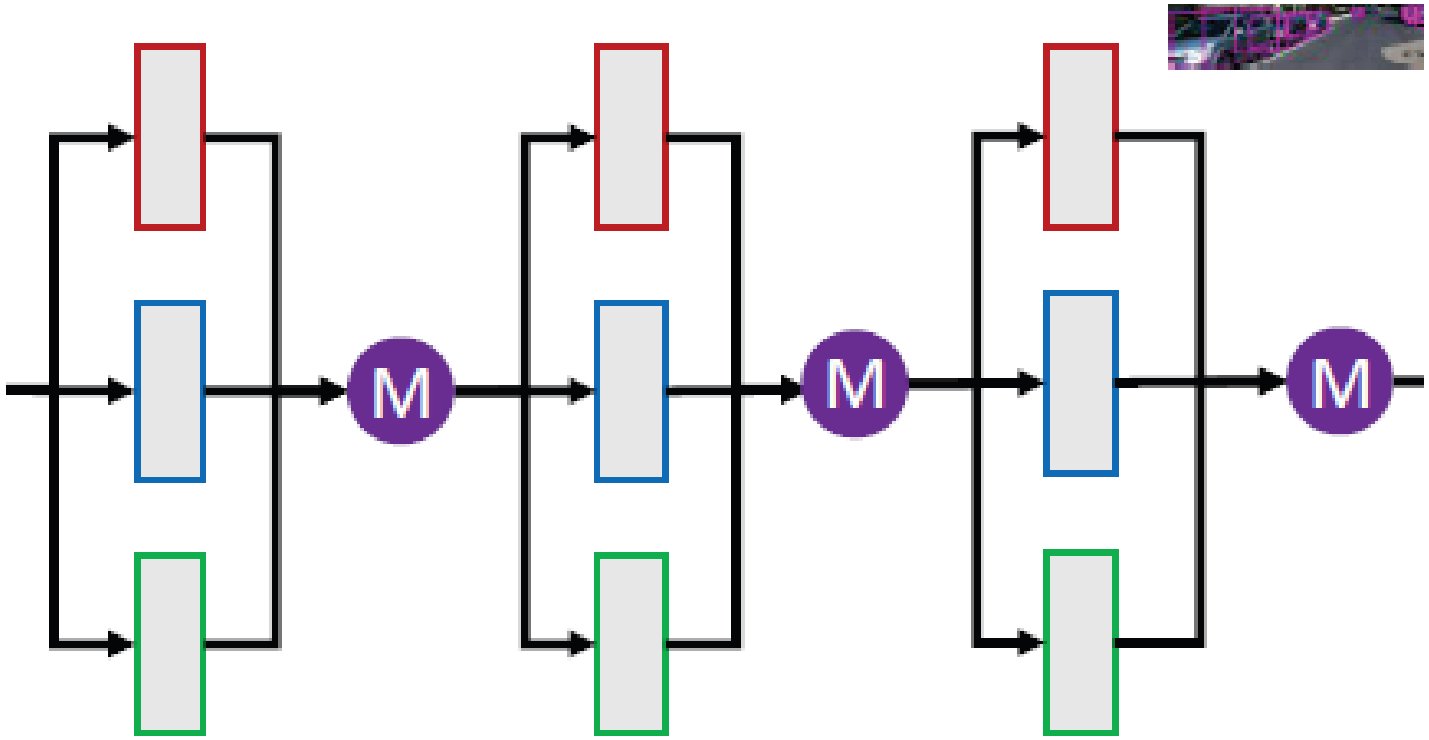
MV3D

Fusion of Feature



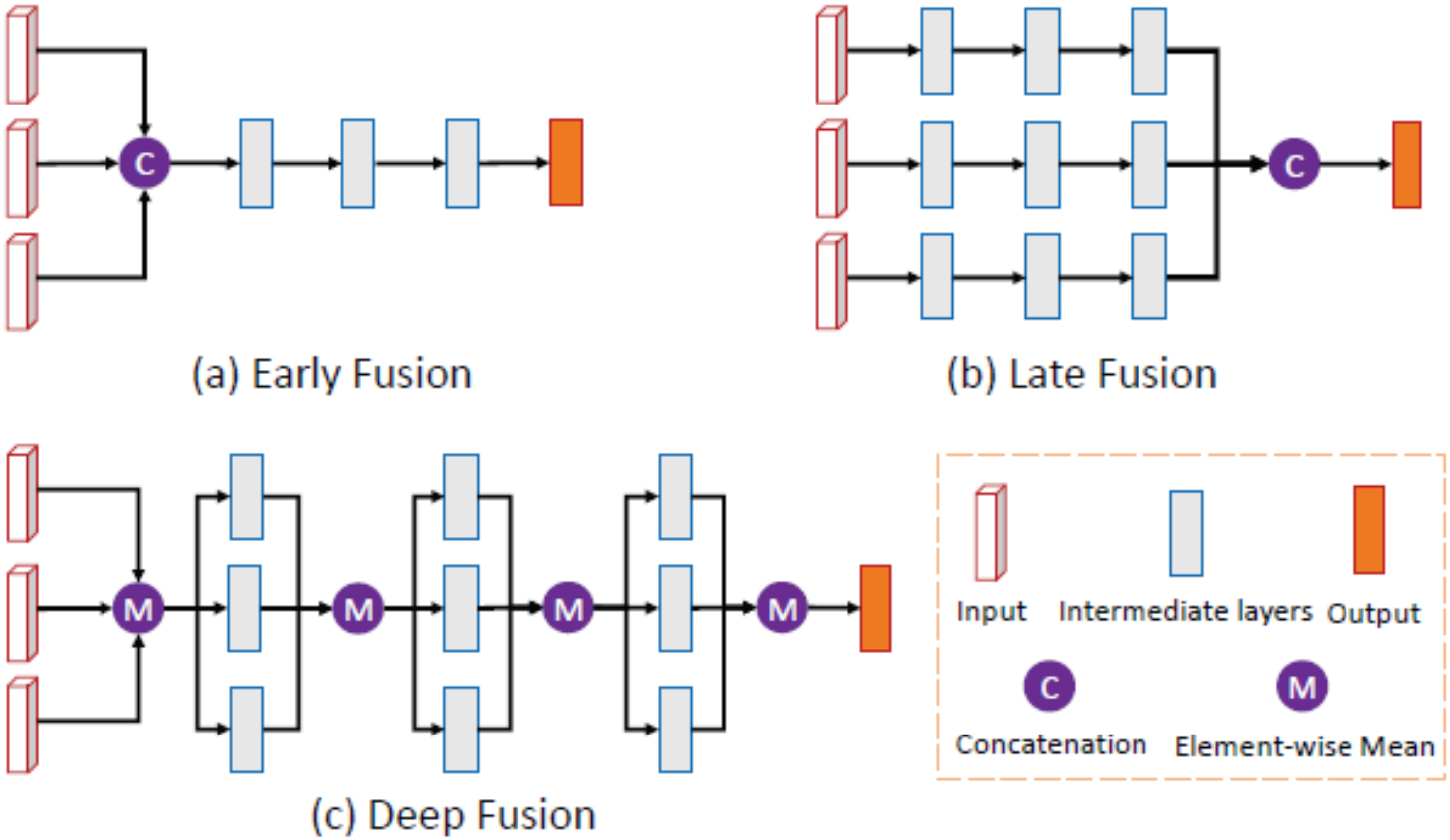
MV3D

Fusion of Feature



MV3D

Fusion of Feature



MV3D

Fusion of Feature

Early Fusion

$$f_L = \mathbf{H}_L(\mathbf{H}_{L-1}(\cdots \mathbf{H}_1(f_{BV} \oplus f_{FV} \oplus f_{RGB})))$$

Late Fusion

$$\begin{aligned} f_L =& (\mathbf{H}_L^{BV}(\cdots \mathbf{H}_1^{BV}(f_{BV})))\oplus \\ & (\mathbf{H}_L^{FV}(\cdots \mathbf{H}_1^{FV}(f_{FV})))\oplus \\ & (\mathbf{H}_L^{RGB}(\cdots \mathbf{H}_1^{RGB}(f_{RGB}))) \end{aligned}$$

Deep Fusion

$$\begin{aligned} f_0 =& f_{BV} \oplus f_{FV} \oplus f_{RGB} \\ f_l =& \mathbf{H}_l^{BV}(f_{l-1}) \oplus \mathbf{H}_l^{FV}(f_{l-1}) \oplus \mathbf{H}_l^{RGB}(f_{l-1}), \\ & \forall l = 1, \cdots, L \end{aligned}$$

MV3D

Comparison

Method	Data	IoU=0.5			IoU=0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D [3]	Mono	30.5	22.39	19.16	5.22	5.19	4.13
3DOP [4]	Stereo	55.04	41.25	34.55	12.63	9.49	7.59
VeloFCN [16]	LIDAR	79.68	63.82	62.80	40.14	32.08	30.47
Ours (BV+FV)	LIDAR	95.74	88.57	88.13	86.18	77.32	76.33
Ours (BV+FV+RGB)	LIDAR+Mono	**96.34**	**89.39**	**88.67**	**86.55**	**78.10**	**76.67**

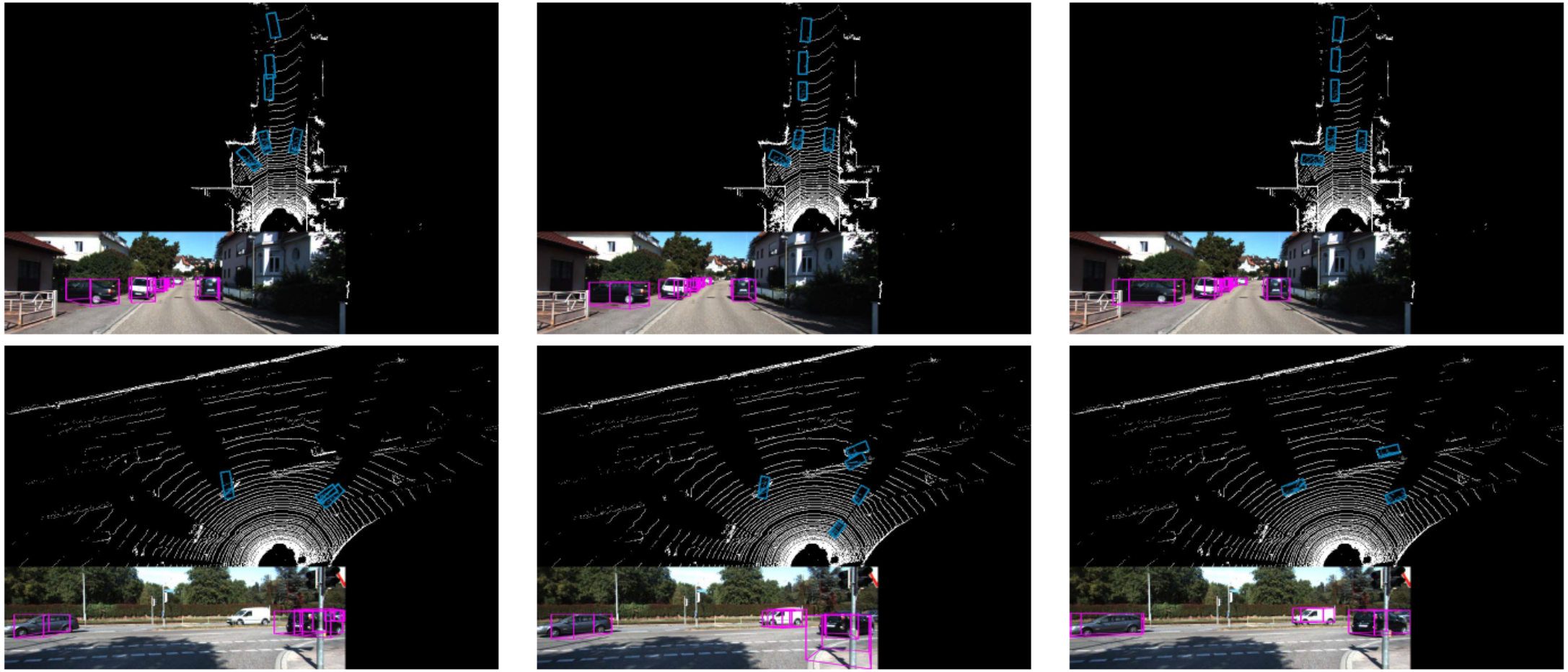
Table 1: **3D localization performance:** Average Precision (AP_{loc}) (in %) of bird's eye view boxes on KITTI *validation* set.

Method	Data	IoU=0.25			IoU=0.5			IoU=0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D [3]	Mono	62.94	48.2	42.68	25.19	18.2	15.52	2.53	2.31	2.31
3DOP [4]	Stereo	85.49	68.82	64.09	46.04	34.63	30.09	6.55	5.07	4.1
VeloFCN [16]	LIDAR	89.04	81.06	75.93	67.92	57.57	52.56	15.20	13.66	15.98
Ours (BV+FV)	LIDAR	96.03	88.85	88.39	95.19	87.65	80.11	71.19	56.60	55.30
Ours (BV+FV+RGB)	LIDAR+Mono	**96.52**	**89.56**	**88.94**	**96.02**	**89.05**	**88.38**	**71.29**	**62.68**	**56.56**

Table 2: **3D detection performance:** Average Precision (AP_{3D}) (in %) of 3D boxes on KITTI *validation* set.

MV3D

Result



3DOP [4] VeloFCN [16] Ours