EBERHARD KARLS UNIVERSITÄT TÜBINGEN

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT

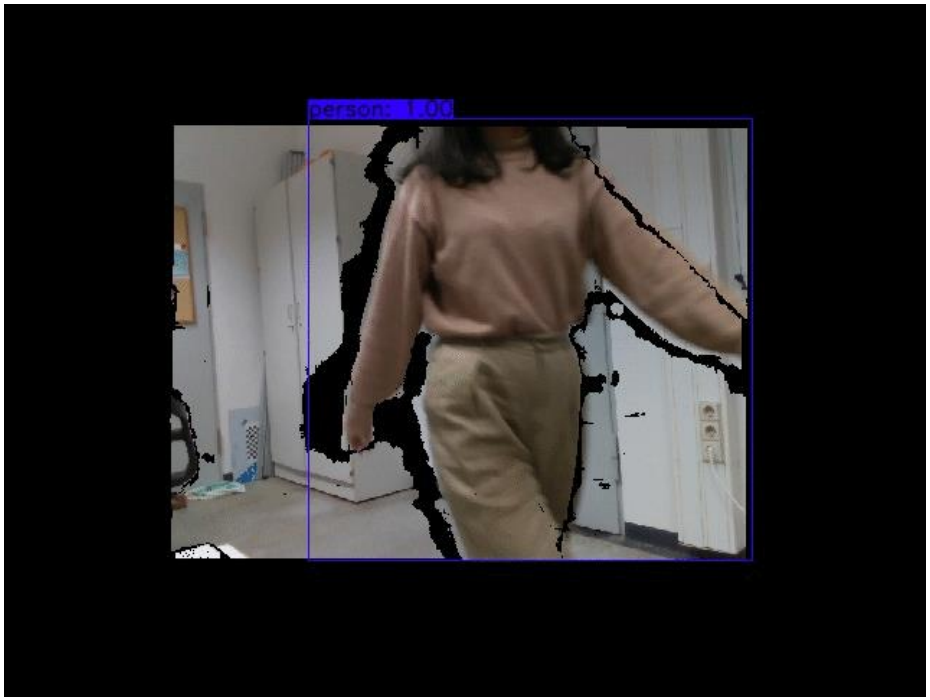
Yolo+FPN: 2D and 3D Fused Object Detection With an RGB-D Camera

Ya Wang, Andreas Zell

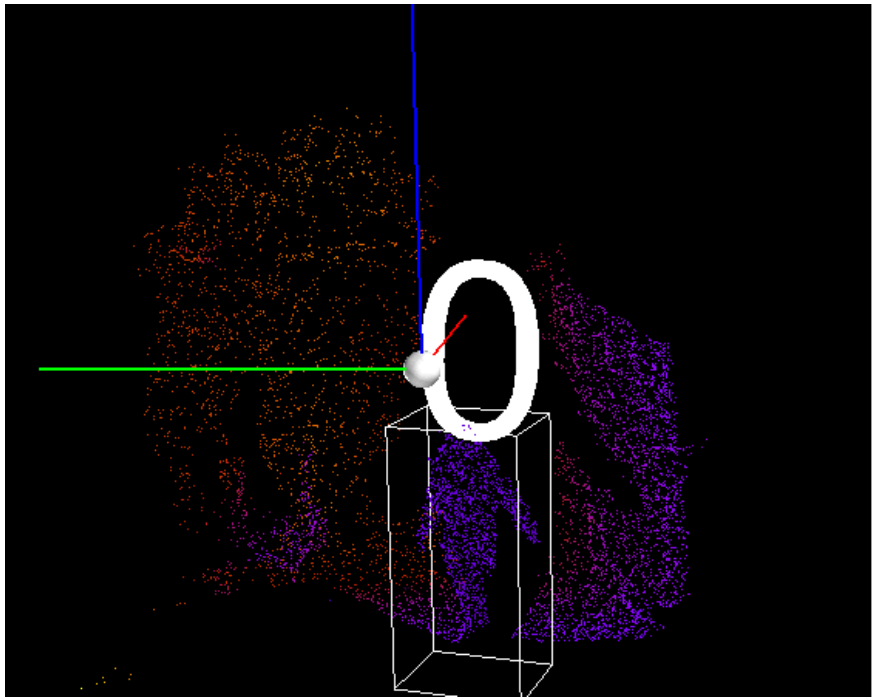
WSI - Cognitive Systems

ICPR 2020

2D and 3D Object Detection



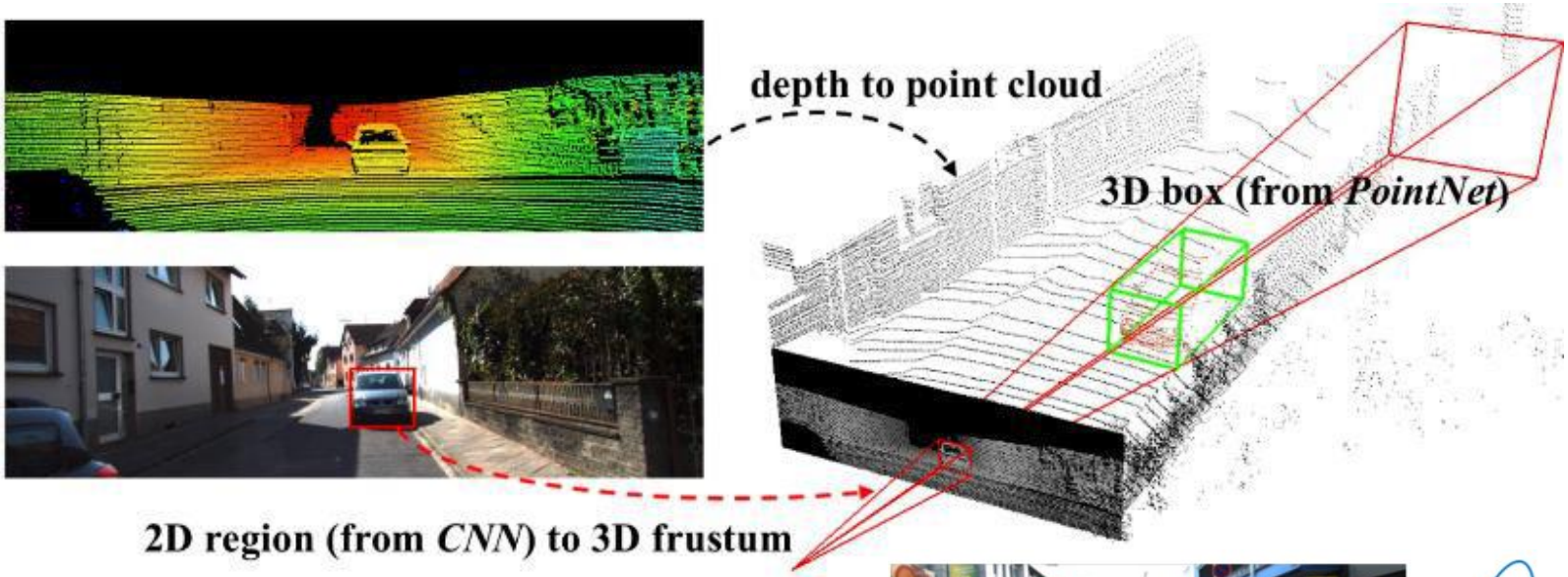
2D Visualization



3D Visualization

How to find optimized fusion strategies for combining 2D with 3D object detection? – **Yolo+FPN**

Introduction and motivation



- Frustum-PointNets:

Lidar with RGB camera (KITTI)

- OurFPN:

1. Real implementation on robots using a single RGB-D camera;

2. Using 2D to help improve 3D object detection

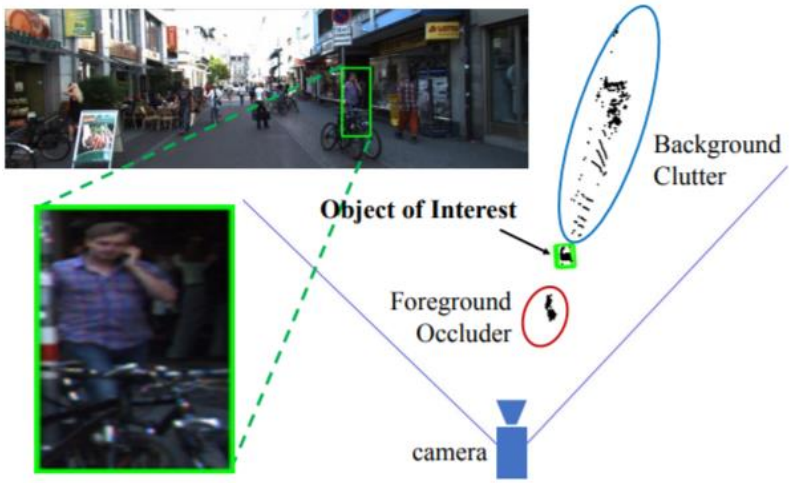


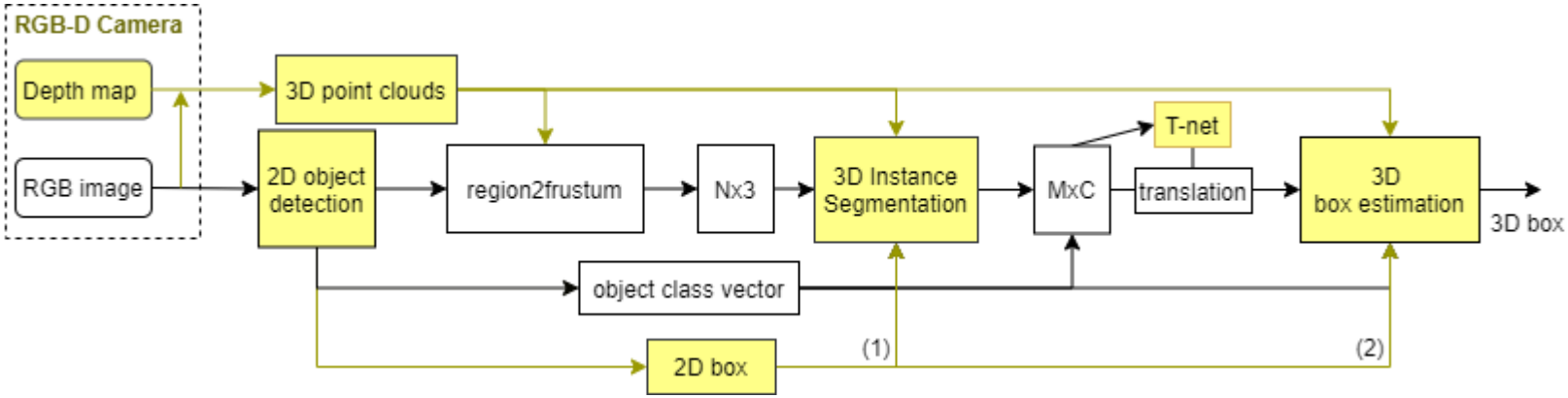
Image source: C. Qi et al., Frustum PointNets, CVPR2018.

Our main difference

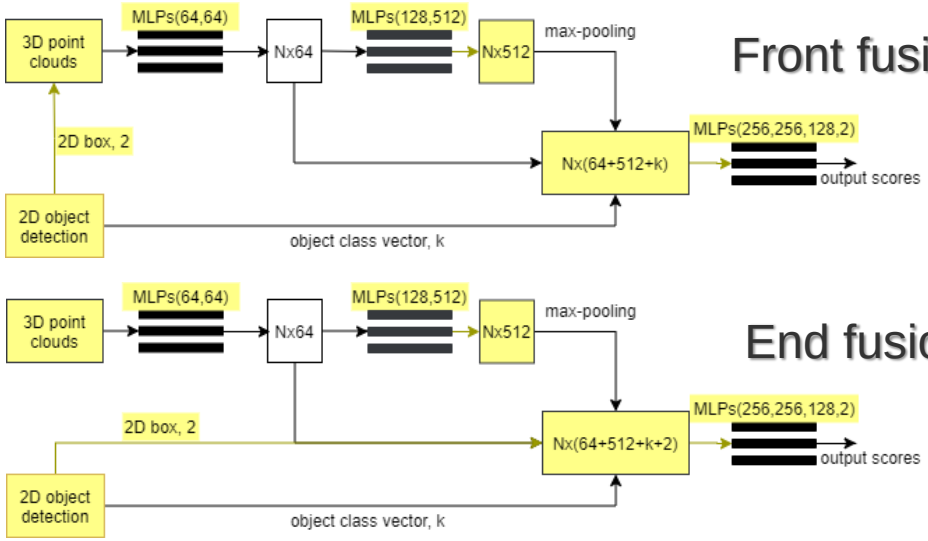
	Frustum PointNets [1]	OurFPN [3]	Yolo+FPN
2D object detection	Faster R-CNN + FPN	YOLOv3, rescale 2D box	Modified YOLOv3, rescale 2D box
3D instance segmentation	PointNet v1 PointNet v2	Reduced PointNet + 2D box object detection	Reduced PointNet + 2D box object detection (different fusion strategies comparison)
Transformation	T-net	T-net	Reduced T-net
3D box estimation	PointNet v1 PointNet v2	Reduced PointNet + 2D box object detection	Reduced PointNet + 2D box object detection (different fusion strategies comparison)

[1] C. R. Qi et al., Frustum PointNets for 3D object detection from RGB-D data, CVPR, pages 918-927, 2018.
[3] Y. Wang et al. Real-time 3D object detection from point clouds using an RGB-D camera. ICPRAM, pages 407-414, 2020.

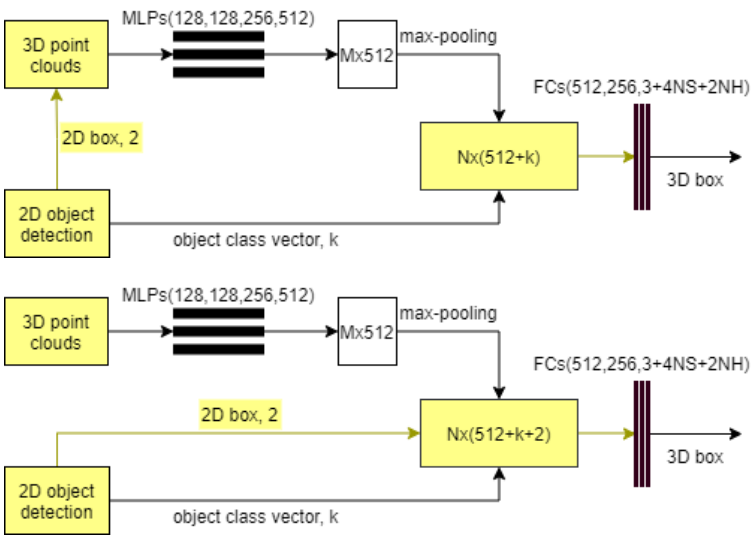
Yolo+FPN



Yolo+FPN Structure



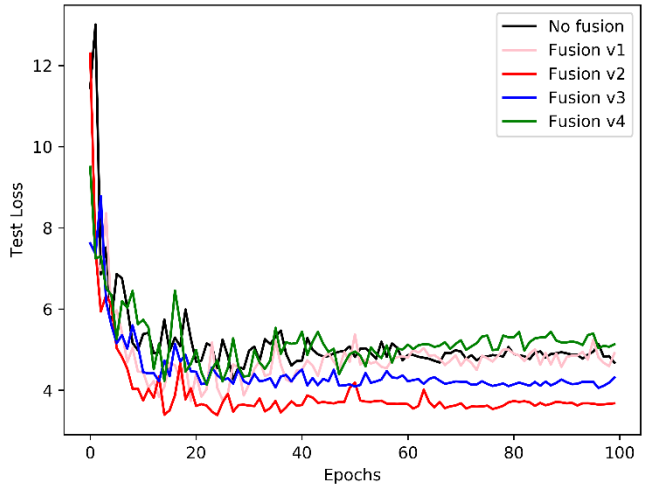
(1) 3D Instance Segmentation



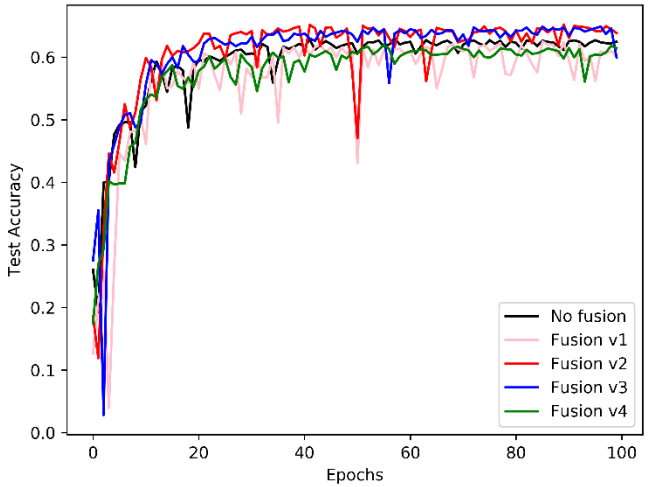
(2) 3D Box Estimation

Quantitative results: KITTI

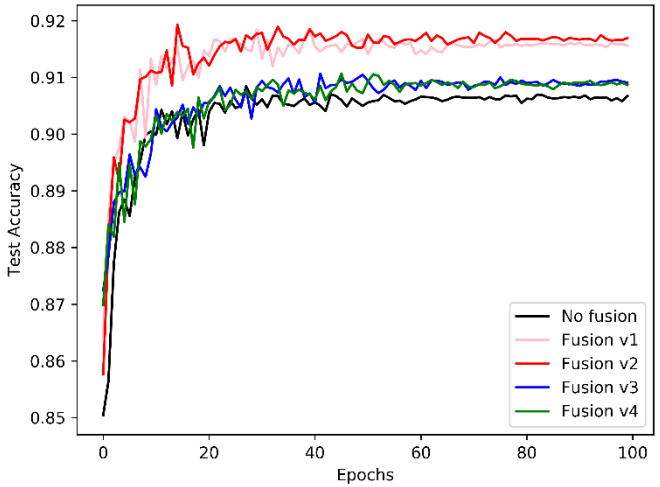
test mean loss:



test box estimation accuracy (IoU=0.7):



test segmentation avg class acc:



- **No fusion:** modified Frustum PointNet
- **Fusion v1:** (1)front fusion + (2)front fusion
- **Fusion v2:** (1)front fusion + (2)end fusion
- **Fusion v3:** (1)end fusion + (2)end fusion
- **Fusion v4:** (1)end fusion + (2)front fusion

Average Precision (AP)

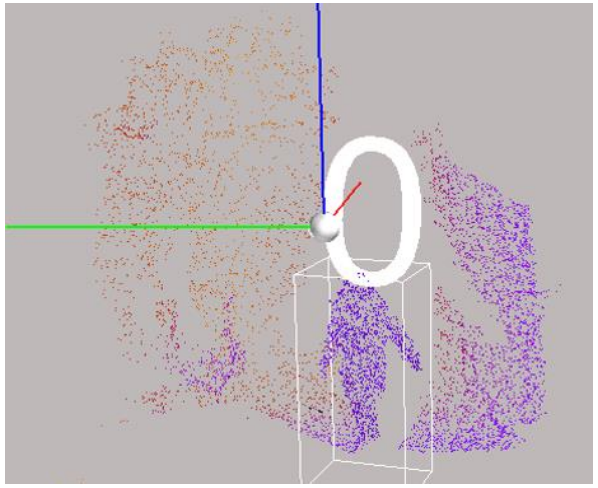
TABLE I: Accuracy Comparison on the KITTI 3D Object Detection Benchmark

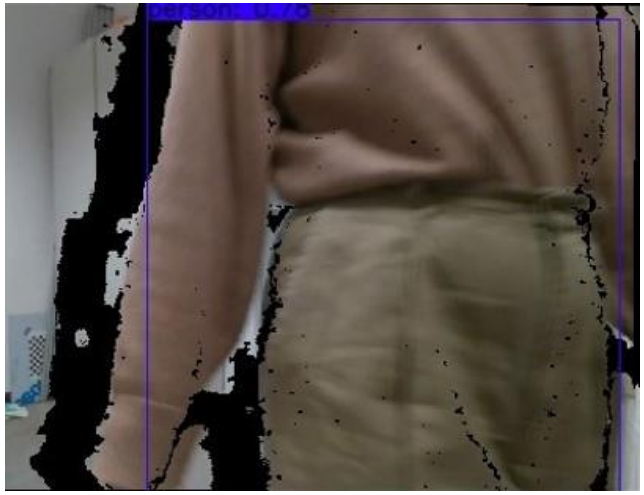
Method	Modality	Pedestrian			Car			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
MV3D [25]	Lidar+RGB	--	--	--	74.97	63.63	54.00	--	--	--
MV3D (Lidar) [25]	Lidar	--	--	--	68.35	54.54	49.16	--	--	--
VoxelNet [1]	Lidar	--	--	--	77.82	64.17	57.51	--	--	--
A-VoxelNet [1]	Lidar	53.66	44.30	40.43	84.01	73.82	66.46	72.58	55.86	49.13
Complexer-YOLO [32]	Lidar	17.60	13.96	12.70	55.93	47.34	42.60	24.27	18.53	17.31
PointRCNN [23]	Lidar	47.98	39.37	36.01	86.96	75.64	**70.70**	74.96	58.82	52.53
Fast PointRCNN [24]	Lidar	--	--	--	85.29	**77.40**	70.24	--	--	--
F-PointNets [2]	Lidar+RGB	50.53	42.15	38.03	82.19	69.79	60.59	72.27	56.12	49.01
F-ConvNet [33]	Lidar+RGB	52.16	43.38	38.80	**87.36**	76.39	66.69	**81.98**	**65.07**	**56.54**
AVOD [26]	Lidar+RGB	36.10	27.86	25.76	76.39	66.47	60.23	57.19	42.08	38.29
AVOD-FPN [26]	Lidar+RGB	50.46	42.27	39.04	83.07	71.76	65.73	63.76	50.55	44.93
Yolo+FPN	RGB-D	**66.83**	**57.68**	**50.94**	85.16	71.86	64.10	76.81	56.94	53.36

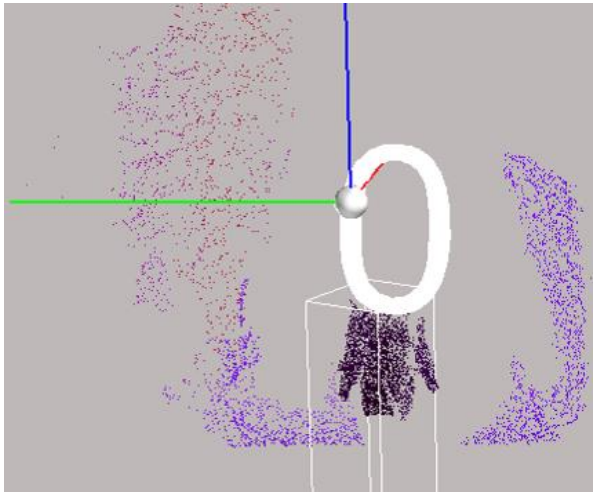
Qualitative results

Global

Local







2D Visualization

3D Visualization

Speed comparison

TABLE II: Speed Comparison of Different Methods

Method	Hardware GPU	FPS
MV3D [25]	GeForce Titan X	2.8
MV3D (Lidar) [25]	GeForce Titan X	4.2
VoxelNet [1]	GeForce Titan X	2.0
Complexer-YOLO [32]	GeForce Titan X	**16.7**
F-PointNets [2]	GeForce GTX 1080	5.9
F-ConvNet [33]	unknown	0.4
AVOD [26]	GeForce Titan Xp	12.5
AVOD-FPN [26]	GeForce Titan Xp	10
Yolo+FPN	Jetson TX2	**10**

FPS:

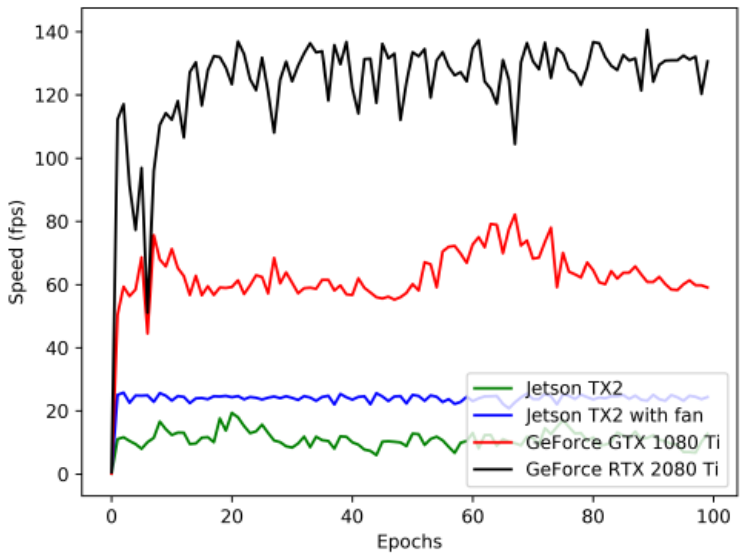


TABLE III: Floating Operations and Parameters Comparison of Detailed Layers

Input Comparison	Frustum PointNet FLOPs	Yolo+FPN FLOPs
2D obj_det	modified Faster R-CNN	modified Yolo v3
3D ins_seg	**910M**	**207M**
T-net	**26M**	**17M**
3D box_est	**93M**	**93M**
Summary	910M+26M+93M=**1029M**	207M+17M+93M=**317M**

Conclusion

- Competitive trade-off between accuracy and speed, compared with some state-of-the-art methods;
- Comparison and analysis of different fusion strategies for 2D and 3D object detection, and can be extended to other methods;
- Real implementation, ready to be used for drones in real-time 2D and 3D object detection with a single RGB-D camera.

Thank you