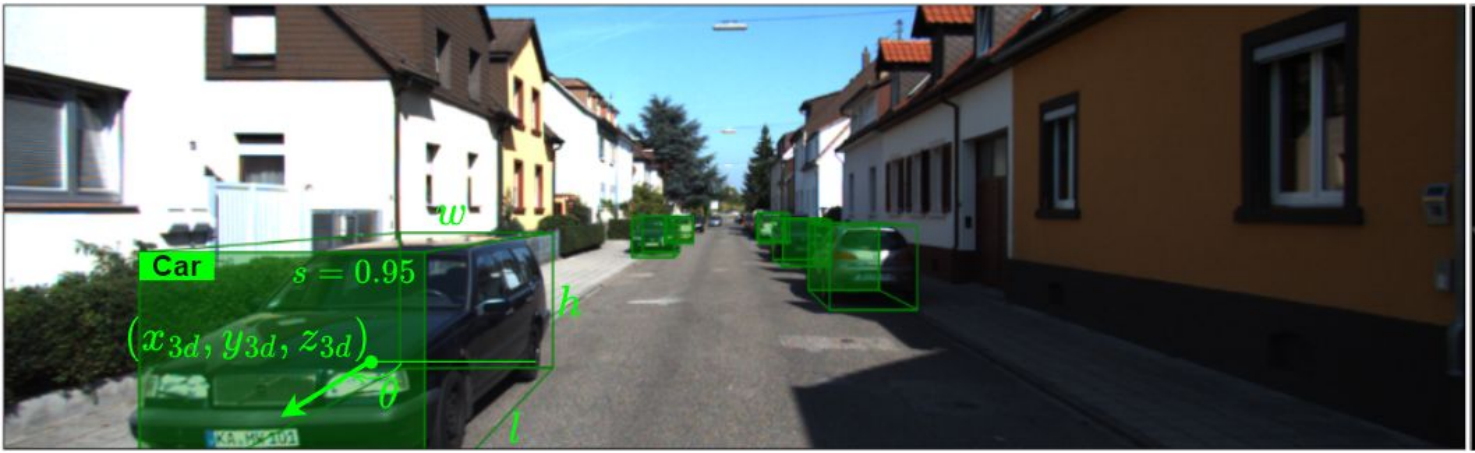
Perspective-aware Convolution for Monocular 3D Object Detection

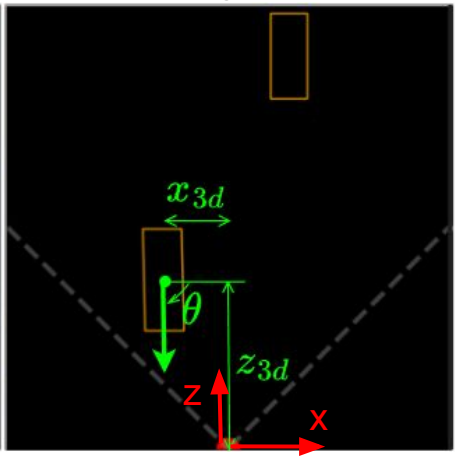
Jia-Quan Yu(游家權), Soo-Chang Pei(貝蘇章)

Introduction - Monocular 3D Object Detection

- We want to find objects location (x_{3d}, y_{3d}, z_{3d}), dimension $(w,\ h,\ l)$, and orientation θ respect to camera center.



Bird's-eye-view

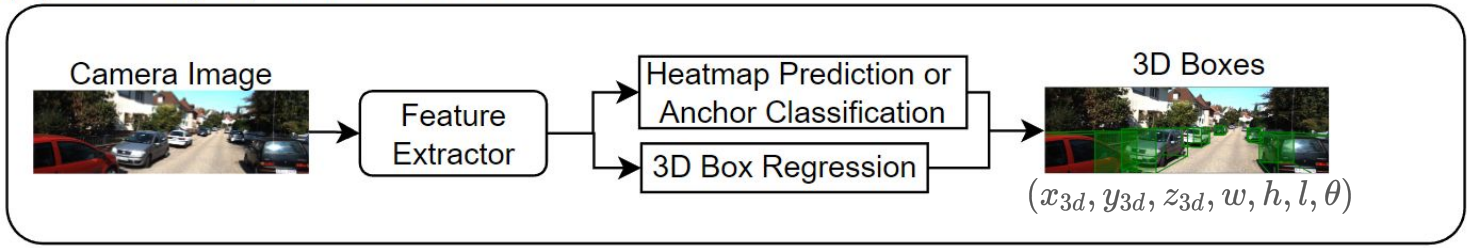


Related Works - Two-stage and One-stage Detectors

(1)Two-stage: based on 2D box prior

- Sensitive to predicted 2D box
- Low Accuracy
- Network: MonoGRNet[2], Deep3Dbox[8]

(2)One-stage: parallel branches



- Faster
- High Accuracy
- Network: SMOKE[3], DD3D[4], MonoFlex[5], Ground-aware[6]

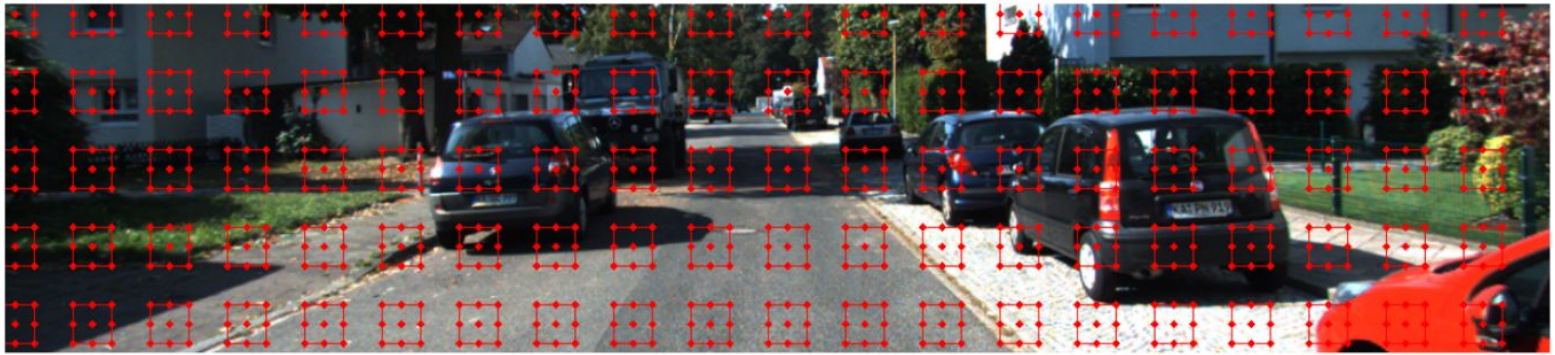
Proposed Method - Perspective-aware Convolution(PAC)

- The idea is to use pictorial clue along the depth-axis to help network to infer object's depth.
- Assumptions:
  - Camera intrinsic matrix is given.
  - Camera height is fixed and the heading is always parallel to the ground.

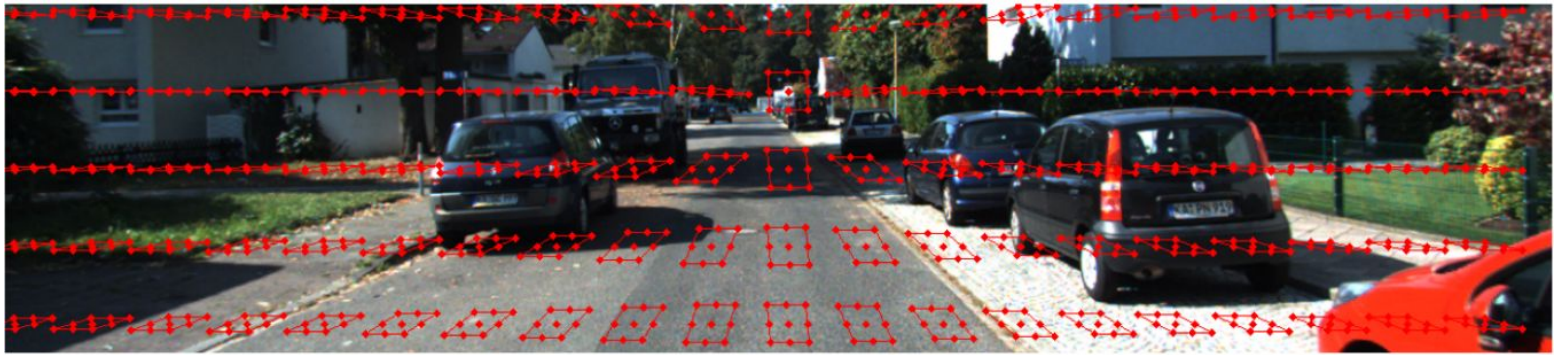
Proposed Method - Perspective-aware Convolution

: Convolutional Kernal

Conventional Convolution

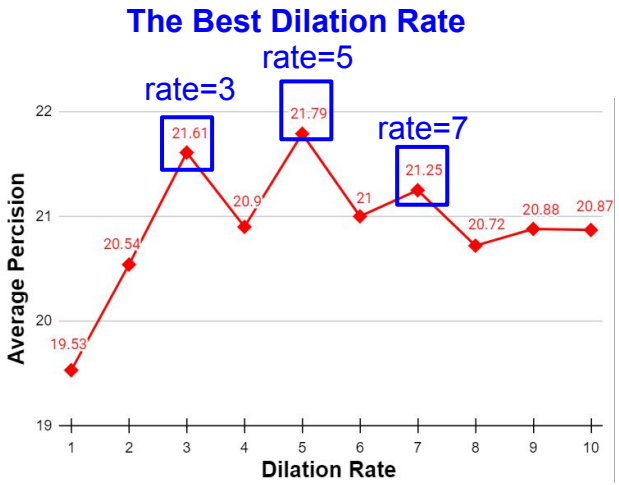


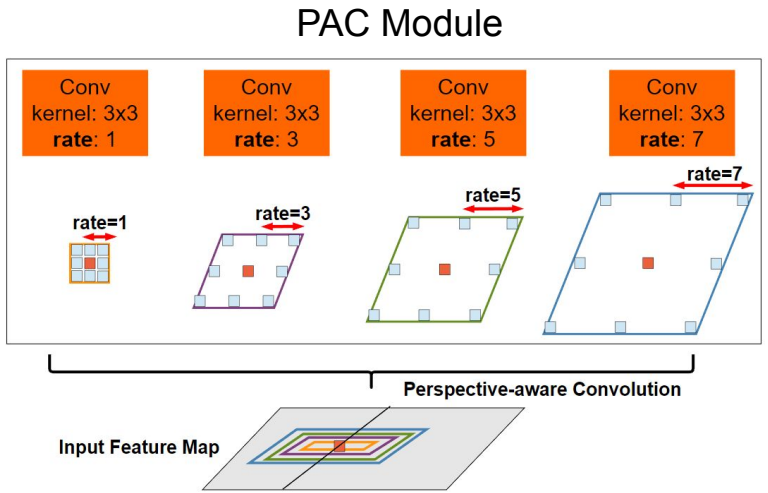
Perspective-aware Convolution(PAC)



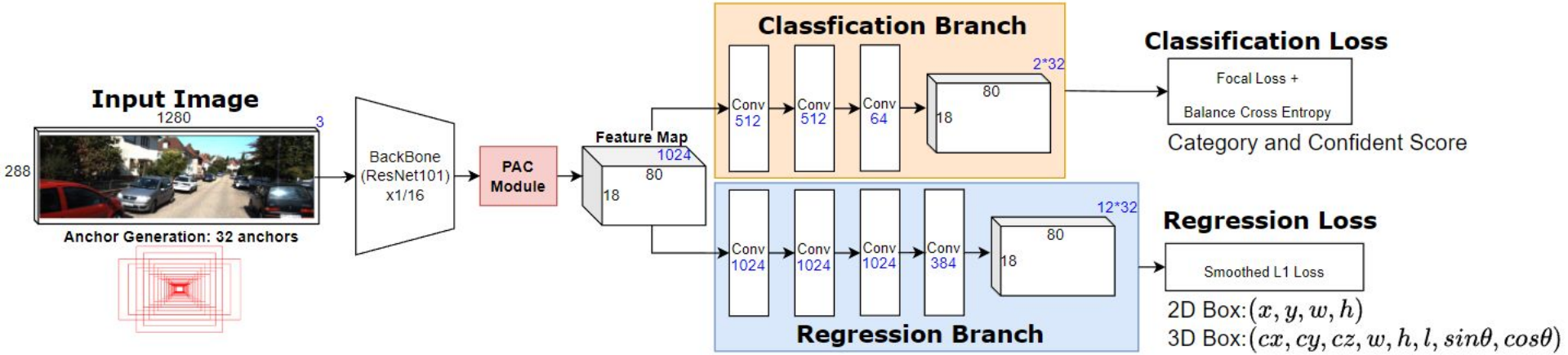
Proposed Method - PAC Module

- Our PAC module incorporates three PAC layers in parallel branches with different dilation rate.





Proposed Architecture - 3D Object Detector with PAC Module



Experiment Result

- Dataset: **KITTI**[7], which include 3,711 training image and 3,768 validation image.
- We train for 30 epochs with batch size of 8.
- We use average precision(AP) as evaluation metric.

Comparison with related works

Methods	Car AP 3D (IoU=0.7)		
	Easy	Moderate	Hard
(1) Two-stage			
MonoGRNet[2]	12.28	7.76	5.91
(2) One-stage			
SMOKE[3]	6.96	4.30	3.98
DD3D[4]	19.16	15.27	13.37
MonoFlex[5]	22.14	16.19	14.18
Ground-aware[6]	21.90	16.06	13.17
Ours (Ground-aware + PAC module)	**23.53**	**17.23**	**14.33**

Ablation Study

Methods	Car AP 3D(IoU=0.7)		
	Easy	Moderate	Hard
Baseline	22.08	15.64	13
PAC(rate=3)	22.59	15.82	12.97
ASPP[1]	22.44	16.96	14.23
PAC Module	**23.53**	**17.23**	**14.33**

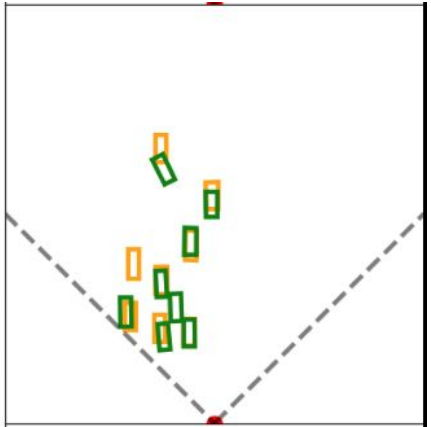
Experiment Result

: Ground Truth

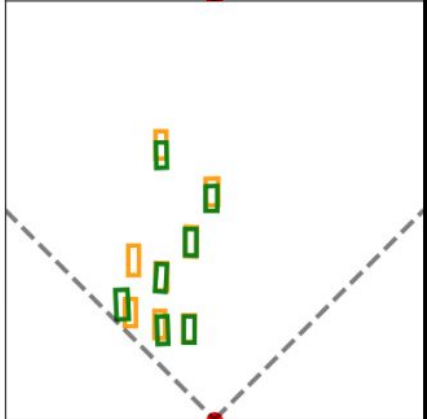
: Prediction

Bird's-eye-view

Without PAC Module



With PAC Module



Experiment Result - Demo Video

Conclusion

- We propose a novel **perspective-aware** convolutional module that able to **adjust kernel shape** based on the camera intrinsic matrix.
- Our experiment result shows image feature along the depth-axis can help to predict object depth.
- Our PAC module **excel in crowded scenes** since it utilizes nearby objects feature to predict object depth.

Reference

- [1] Rethinking Atrous Convolution for Semantic Image Segmentation, https://arxiv.org/abs/1706.05587
- [2] MonoGRNet: A Geometric Reasoning Network for Monocular 3D Object Localization, https://arxiv.org/abs/1811.10247
- [3] SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation, https://arxiv.org/abs/2002.10111
- [4] Is Pseudo-Lidar needed for Monocular 3D Object detection?, https://arxiv.org/abs/2108.06417
- [5] Objects are Different: Flexible Monocular 3D Object Detection, https://arxiv.org/abs/2104.02323
- [6] Ground-aware Monocular 3D Object Detection for Autonomous Driving, https://arxiv.org/abs/2102.00690
- [7] KITTI dataset, https://www.cvlibs.net/datasets/kitti/
- [8] 3D Bounding Box Estimation Using Deep Learning and Geometry, https://arxiv.org/abs/1612.00496
- [9] Our code on github, https://github.com/KenYu910645/perspective-aware-convolution