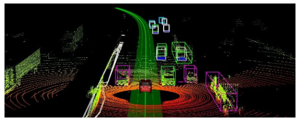
Center Based 3D Object Detection and Tracking with BiFPN Blocks

*Dan Halperin** [1], *Koray Koca** [1]

**Both authors contributed equally,* [1]*Technical University of Munich*

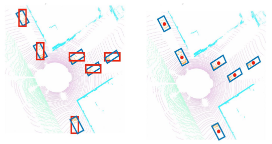
Motivation

- LIDAR scanners have many advantages, when used in practice.
- Detecting 3D objects within point clouds is specifically crucial for autonomous driving.

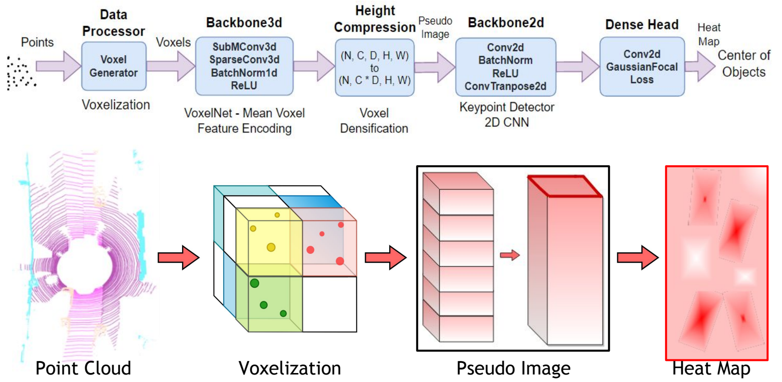


Baseline - CenterPoint

- 3D outdoor objects detection, that is robust to alignment with the axes.



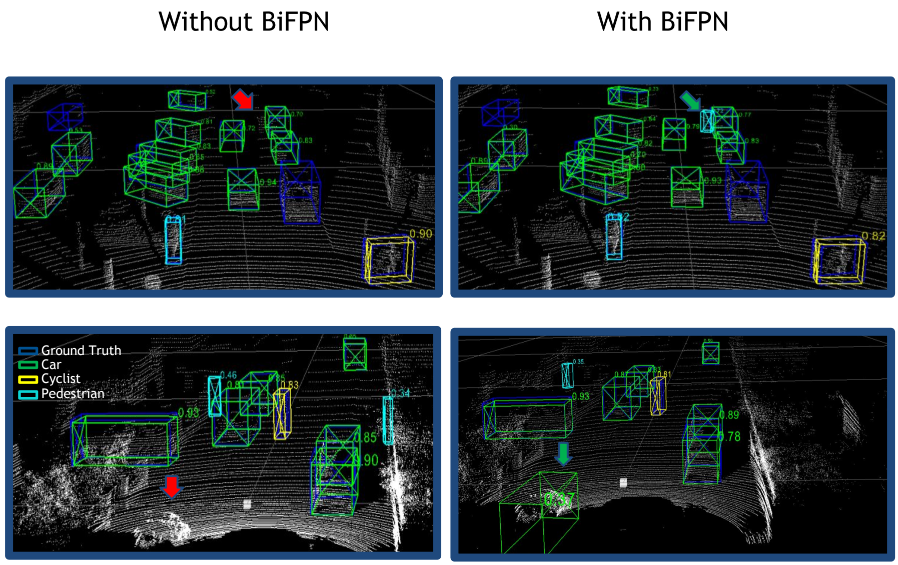
Overview - Baseline Architecture



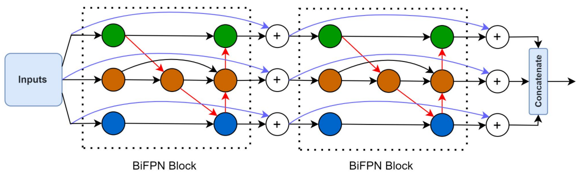
Datasets

Dataset	Enviroment	Scans	Scenes	Split*	Classes	Task	Metric
KITTI	Karlsruhe	14,999	400	0.5/0.25/0.25	11	3D detection	mAP
nuScenes	Boston, Singapore	390,000	1000	0.7/0.15/0.15	23	3D detection and tracking	mAP

Examples



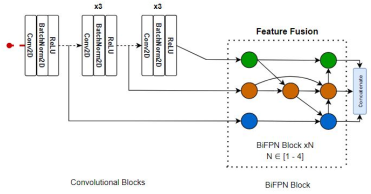
BiFPN Blocks with skip-connections



Nodes : Conv → BN → ReLU, **Arrows** Weighted Fusion

Overview of the proposed architecture: The upper figure shows the BiFPN blocks that we integrated inside the 2D-backbone CNN network. Nodes represent feature maps. The main advantages of this block are:

- Top-Down and Bottom-Up Multiscale Feature Fusion
- Skip connections to control and learn the flow of information to deeper layers and to overcome optimization problems like vanishing/exploding gradients



Inputs of the BiFPN block: Inputs are taken from the different feature levels throughout the 2D CNN Network. Hence, the high level and low level features are fused. This block can be used the number of N times repetitively.

Hardware

For training and testing, we used a virtual machine in GCP with the following settings:

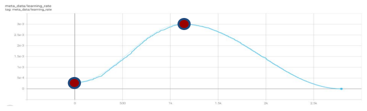
- Intel N1-highmem-4 CPU (4vCPU, 26 GB Memory)
- Nvidia Tesla T4 16 GB GPU
- 500 GB SSD

Quantitative Results

model	Recall 0.3	Recall 0.5	Recall 0.7	AP: Cars	AP: Pedestrians	AP: Cyclists	#Parameters (M)
Originial	95.03%	88.44%	65.66%	**96.82%**	70.42%	**92.99%**	**1.880**
BiFPN: [128]	93.67%	87.74%	62.73%	96.27%	52.85%	75.03%	2.913
BiFPN: [128] + SC	**95.44%**	**89.23%**	66.10%	95.31%	**73.02%**	91.60%	2.963
BiFPN: [128, 64] + SC	95.33%	89.16%	**66.68%**	96.45%	**73.02%**	90.24%	3.346

Interesting Parameters

Cyclic learning rate



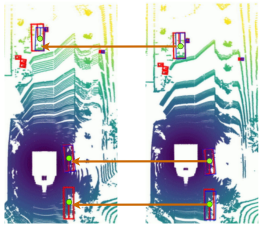
Gaussian focal loss

$$\mathrm{CE}(p_t) = -\log(p_t)$$

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$$

Tracking

For tracking, the network project the object centers in the current frame back to the previous frame, and then matches them to the tracked objects by closest distance matching.



Conclusion

- Better generalization for rare classes like cyclists and pedestrians on KITTI Dataset.
- Less number of False Positives.
- With a better hardware and optimization, we expect to replicate the improvement in detection results, as observed in [3].
- A loss function that is more robust to smaller objects could improve performance for all classes.
- A decaying learning rate might have performed better - small batch size already introduces regularization.

Future work: Training on more epochs with full NuScenes Dataset

Main References

[1] T. Yin et al., “Center-based 3d object detection and tracking”, In *CVPR*, 2021
[2] M. Tan et al., “Efficientdet: Scalable and efficient object detection, In *CVPR*, 2020
[3] Z. Ding et al., “1st Place Solution for Waymo Open Dataset Challenge - 3D Detection and Domain Adaptation, 2020