FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection




Tai Wang, Xinge Zhu, Jiangmiao Pang, Dahua Lin

The Chinese University of Hong Kong

{wt019, zx018, dhlin}@ie.cuhk.edu.hk, pangjiangmiao@gmail.com

1. Introduction
2. Related Work
3. Approach
4. Experiments
5. Follow-ups

Introduction

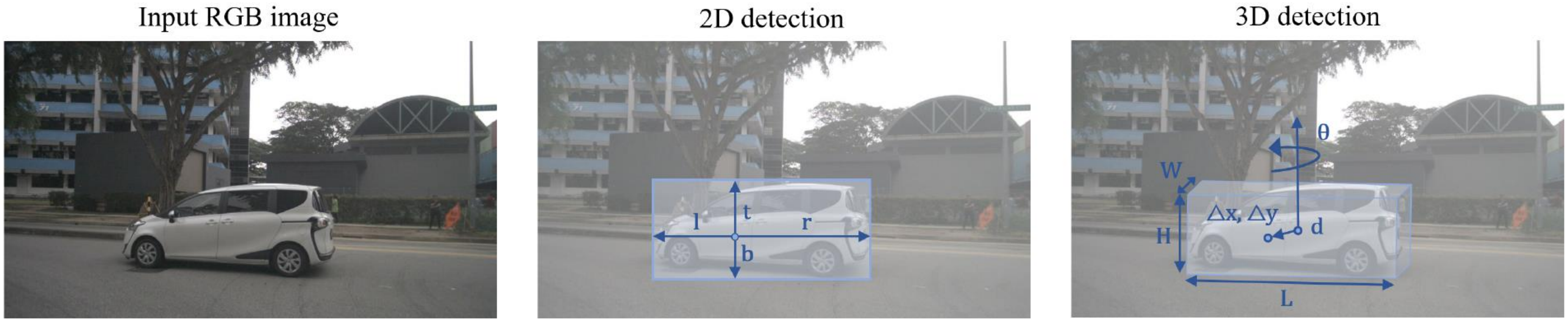
- **Background**
  - Well-developed 2D detection & LiDAR-based 3D detection
  - The performance of monocular 3D detection still lags far behind
  - ➡ Revisit monocular 3D detection from a higher level:
    - Unified detection paradigms/modules
    - Generalized methodology for transferring successful experiences (across different settings/metrics/detectors)
    - A simple yet effective and efficient baseline

Introduction

- **Our Approach – Study how to adapt a 2D detector for 3D detection**
  - Transform 7-DoF 3D targets to the image domain
  - A practice built on FCOS
    - Distribute objects according to 2D scales
    - Assign targets according to the projected 3D-center
    - Re-define the center-ness with a 2D Gaussian distribution
  - A simple yet effective detector

Related Work

- **2D Detection**
  - Anchor-based vs. anchor-free (more suitable for monocular 3D detection)
  - Closely related to monocular 3D detection but the connection is usually ignored

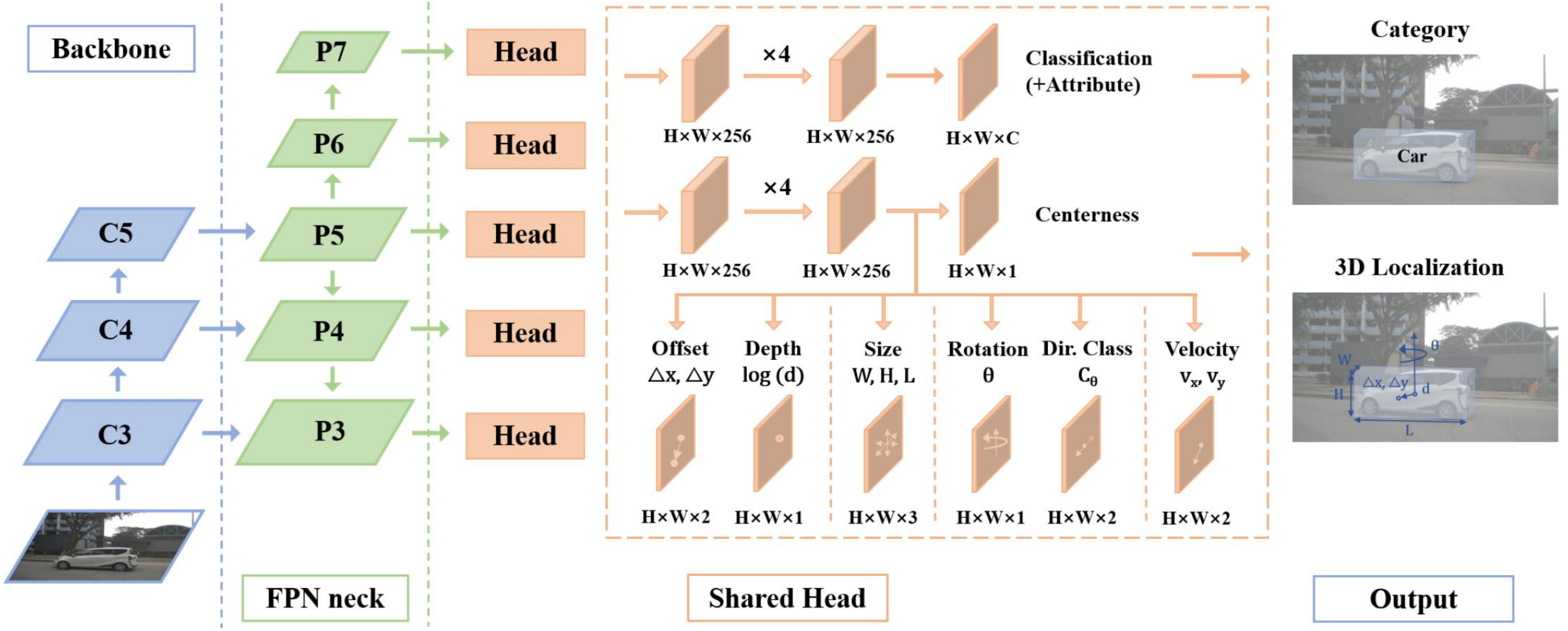


Related Work

- **Monocular 3D Detection**
  - Methods involving sub-networks (3DOP[1], MLFusion[2], Deep3DBox[3])
    - Rely on the performance of sub-networks, external data and pre-trained models
  - Transform to 3D representations (Pseudo-LiDAR[4], PatchNet[5], OFTNet[6])
    - Rely on dense depth labels
    - Involve domain gaps between different depth sensors
  - End-to-end design like 2D detection (M3D-RPN[7], SS3D[8], MonoDIS[9], RTM3D[10])
    - Lacks unified and generalized designs
  - Few works study the key difficulty when applying a 2D detector on this 3D task

Approach

- **Framework Overview**

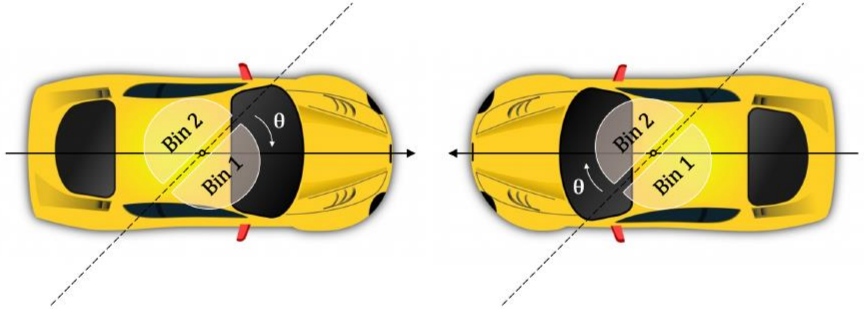


Approach

▶ Framework Overview

- Backbone and FPN neck following FCOS [11]
- Detection Head: classification & localization
  - Regression targets:

    $\Delta x, \Delta y$ (2D attributes); $\log(d), W, H, L, \theta, C_\theta, v_x, v_y$ (3D attributes)



- Loss design:

$$L_{cls} = -\alpha(1-p)^\gamma \log p \qquad L_{loc} = \sum_{b_i \in (\Delta x, \Delta y, d, W, H, L, sin\theta, v_x, v_y)} \omega_i \, SmoothL1(\Delta b_i)$$

Other classification losses: $L_{attr}/L_{dir}/L_{ct}$

$$L = \frac{1}{N_{pos}}(\beta_{cls}L_{cls} + \beta_{loc}L_{loc} + \beta_{attr}L_{attr} + \beta_{dir}L_{dir} + \beta_{ct}L_{ct}), \text{ all } \beta = 1.0$$

Approach

- **2D Guided Multi-Level 3D Prediction**
  - Distribute objects according to 2D scales
    - 2D regression targets → distribute objects
    - Criterion: $m_{i-1} < \max(l^*, r^*, t^*, b^*) < m_i,\ m \in (0, 48, 96, 192, 384, \infty)$
  - Assign targets based on projected 3D-centers
    - Center-sampling strategy → 3D-center
    - Ambiguity problem: A fore-ground point corresponds to multiple targets
      - Adopt the distance priority principle instead of area priority

        (Improve the best possible recall (BPR) and mAP for large objects)

Approach

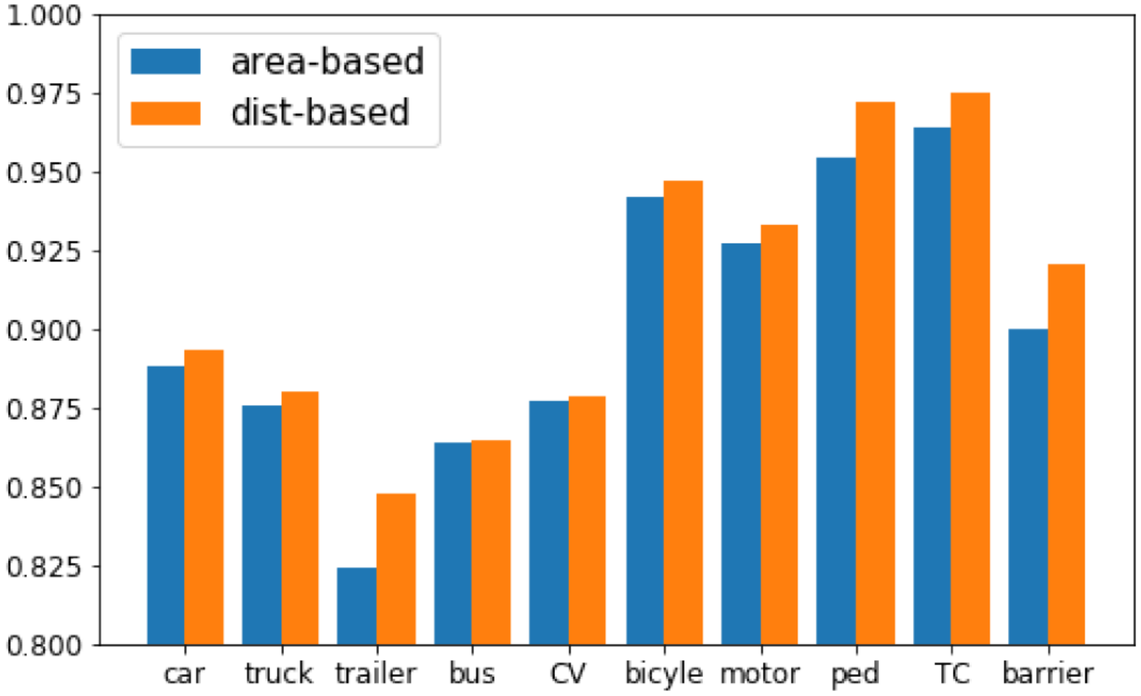


Figure 4: Our proposed distance-based target assignment for dealing with ambiguity case could significantly improve the best possible recall (BPR) for each class, especially for large objects like trailers. Construction vehicle and traffic cone are abbreviated as CV and TC in this figure.

Approach

- **3D Center-ness with 2D Gaussian Distribution**
  - 2D center-ness in FCOS [11] :

$$c = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}$$

  - 3D center-ness in FCOS3D:

$$c = e^{-\alpha((\Delta x)^2+(\Delta y)^2)}, \quad \alpha = 2.5 \text{ in the experiments.}$$

  - Also use this 3D center-ness to filter low-quality predictions

Experiments

- **Dataset – NuScenes Dataset** [12]
  - Multi-modal data, 700/150/150 scenes for train/val/test
  - RGB images from 6 surround-view cameras
  - 1.4M annotated 3D bounding boxes, 10 categories
- **Evaluation Metrics – NuScenes Detection Score (NDS)**
  - More comprehensive, more tolerant to not strictly precise detections
  - Average Precision metric: $mAP = \frac{1}{|\mathbb{C}||\mathbb{D}|}\sum_{c\in\mathbb{C}}\sum_{d\in\mathbb{D}} AP_{c,d}, \mathbb{D} = \{0.5, 1, 2, 4\}$
  - True Positive metric: $mTP = \frac{1}{|\mathbb{C}|}\sum_{c\in\mathbb{C}} TP_c$ (5 TP metrics: ATE/ASE/AOE/AVE/AAE)
  - NuScenes Detection Score: $NDS = \frac{1}{10}[5mAP + \sum_{mTP\in\mathbb{TP}}(1 - \min(1, mTP))]$

Experiments

- **Implementation Details**
  - Architecture:

    ResNet 101 (Pretrained on ImageNet) + DCN + FPN based on MMDetection3D [13]
  - Training Parameters:

    SGD, batch size 16 on 8 GPUs
  - Finetuning for more competitive performance:

    depth weight = 0.2 (12 epochs) $\rightarrow$ 1.0 (12 epochs)
  - Data Augmentation: only image flip

Experiments

Results

Table 1: Results on the nuScenes dataset.

Methods	Dataset	Modality	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
CenterFusion [22]	test	Camera & Radar	0.326	0.631	0.261	0.516	0.614	0.115	0.449
PointPillars [14]	test	LiDAR	0.305	0.517	0.290	0.500	0.316	0.368	0.453
MEGVII [40]	test	LiDAR	**0.528**	0.300	0.247	0.379	0.245	0.140	**0.633**
LRM0	test	Camera	0.294	0.752	0.265	0.603	1.582	0.14	0.371
MonoDIS [30]	test	Camera	0.304	0.738	0.263	0.546	1.553	0.134	0.384
CenterNet [38] (HGLS)	test	Camera	0.338	0.658	0.255	0.629	1.629	0.142	0.4
Noah CV Lab	test	Camera	0.331	0.660	0.262	0.354	1.663	0.198	0.418
FCOS3D (Ours)	test	Camera	**0.358**	0.690	0.249	0.452	1.434	0.124	**0.428**
CenterNet [38] (DLA)	val	Camera	0.306	0.716	0.264	0.609	1.426	0.658	0.328
FCOS3D (Ours)	val	Camera	**0.343**	0.725	0.263	0.422	1.292	0.153	**0.415**

Experiments

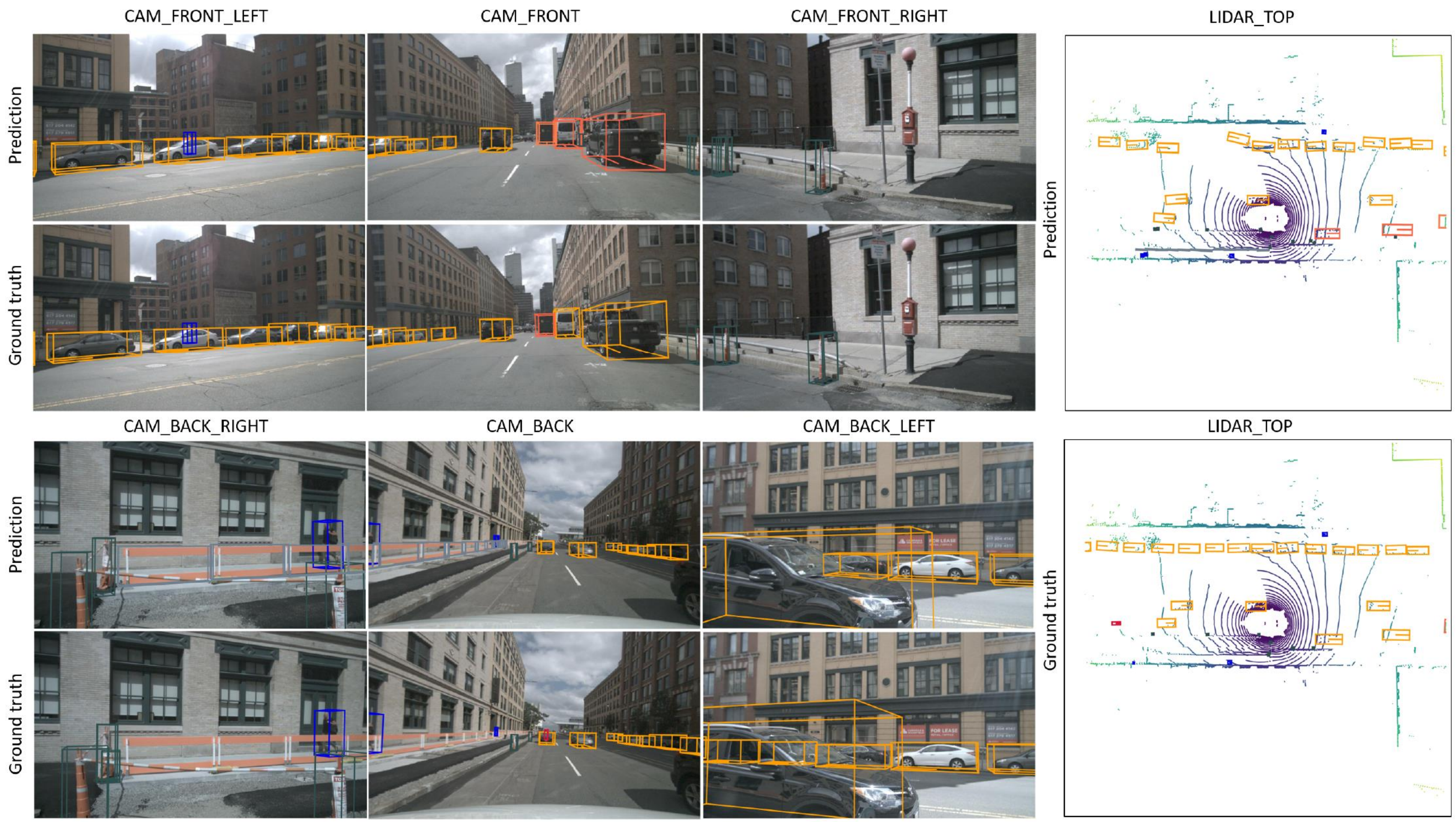
- **How to push it towards SOTA…**

Table 3: Ablation studies on the nuScenes validation 3D detection benchmark.

Methods	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
Baseline (FCOS + 3D targets)	0.227	0.868	0.272	0.778	1.326	0.393	0.282
+ Depth loss in original space	0.25	0.838	0.268	0.892	1.33	0.413	0.284
+ Flip augmentation	0.248	0.85	0.267	1.016	1.358	0.268	0.286
+ Dist-based target assign & attr pred	0.257	0.832	0.268	0.852	1.2	0.18	0.316
+ NMS among predictions of six views	0.26	0.828	0.267	0.85	1.371	0.18	0.317
+ Stronger backbone (ResNet101)	0.272	0.821	0.265	0.81	1.379	0.17	0.329
+ Disentangled heads	0.28	0.822	0.274	0.64	1.305	0.177	0.349
+ DCN in backbone	0.295	0.806	0.268	0.511	1.315	0.17	0.372
+ Finetune w/ depth weight=1.0	0.316	0.755	0.263	0.458	1.307	0.169	0.393
+ Test time augmentation	0.326	0.743	0.259	0.441	1.341	0.163	0.402
+ More epochs & ensemble	**0.343**	0.725	0.263	0.422	1.292	0.153	**0.415**

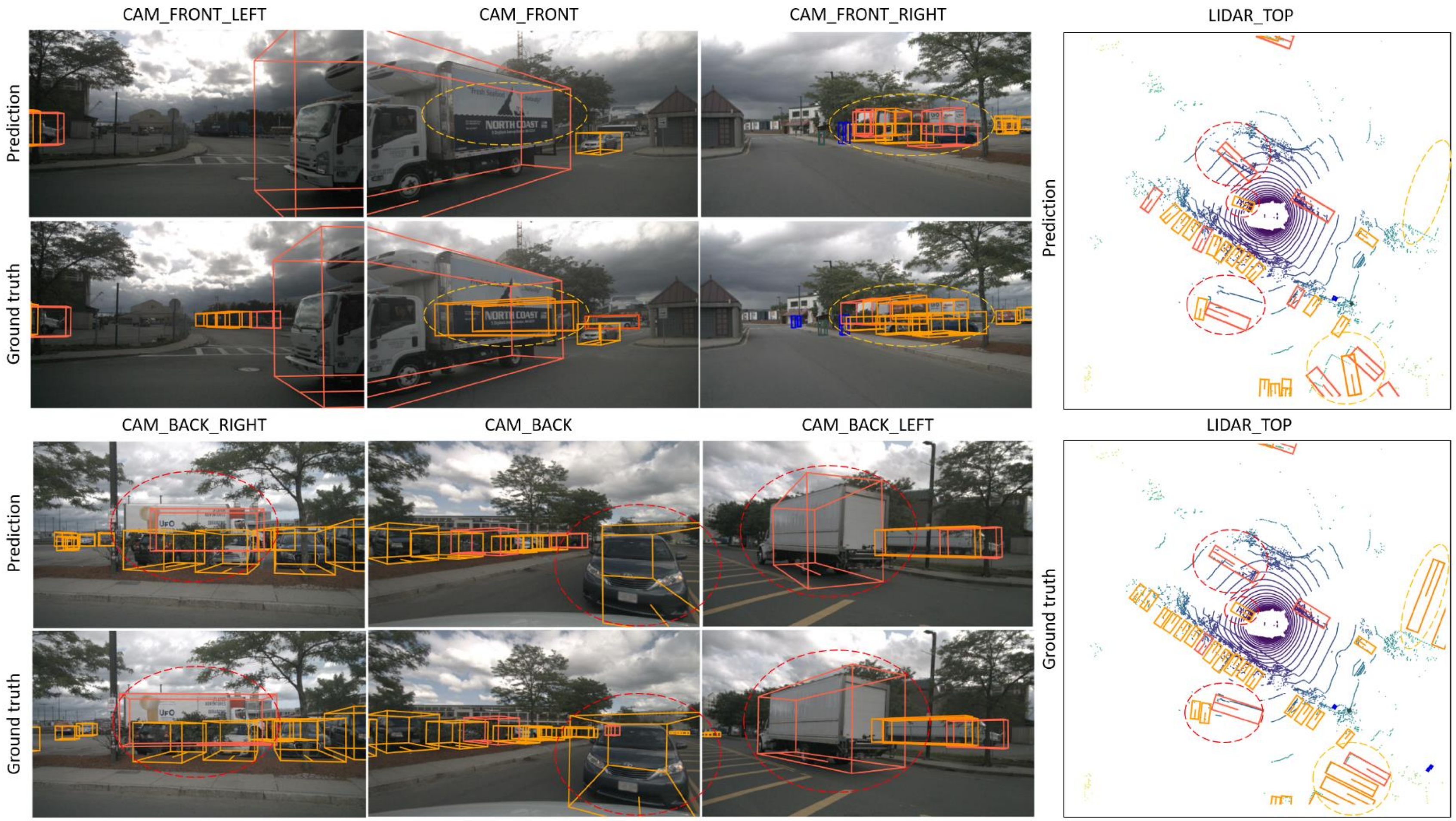
Experiments

- **Qualitative Results**



Experiments

- **Failure Cases**



Follow-ups

- **What's next after unified paradigms?**
  - Generalize them across datasets & What is the key challenge?
    - Probabilistic and Geometric Depth (PGD) [14], CoRL 2021
    - Current monocular 3D detection → instance depth estimation
    - Quite different performance under different settings/metrics
  - Borrow ideas from 2D & connection with 2D
    - Module design of detectors: DETR3D [15]
    - More connections: pretraining in Mono3D → DD3D [16]
  - General multi-view settings: DETR3D [15], ImVoxelNet [17]

Reference

[1] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In Conference on Neural Information Processing Systems, 2015.

[2] B. Xu and Z. Chen. Multi-level fusion based 3d object detection from monocular images. In IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[3] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bounding box estimation using deep learning and geometry. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[4] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q.Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[5] X. Ma, S. Liu, Z. Xia, H. Zhang, X. Zeng, and W. Ouyang. Rethinking Pseudo-LiDAR Representation. In European Conference on Computer Vision (ECCV), 2020.

[6] T. Roddick, A. Kendall, and R. Cipolla. Orthographic feature transform for monocular 3d object detection. CoRR, abs/1811.08188, 2018. URL https://arxiv.org/abs/1811.08188.

[7] G. Brazil and X. Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In IEEE International Conference on Computer Vision, 2019.

[8] E. J¨orgensen, C. Zach, and F. Kahl. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. CoRR, abs/1906.08070, 2019. URL https: //arxiv.org/abs/1906.08070.

[9] A. Simonelli, S. R. R. Bul`o, L. Porzi, M. L´opez-Antequera, and P. Kontschieder. Disentangling monocular 3d object detection. In IEEE International Conference on Computer Vision, 2019.

Reference

[10] P. Li, H. Zhao, P. Liu, and F. Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In European Conference on Computer Vision, 2020.

[11] Z. Tian, C. Shen, H. Chen, and T. He. Fcos: Fully convolutional one-stage object detection. In IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[12] 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. CoRR, abs/1903.11027, 2019. URL http://arxiv.org/abs/1903.11027.

[13] M. Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[14] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In Conference on Robot Learning (CoRL), 2021

[15] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In Conference on Robot Learning (CoRL), 2021

[16] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li and Adrien Gaidon. Is Pseudo-Lidar needed for Monocular 3D Object detection? In IEEE/CVF International Conference on Computer Vision (ICCV), 2021

[17] Danila Rukhovich, Anna Vorontsova and Anton Konushin. ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection. In IEEE Winter Conference on Applications of Computer Vision (WACV), 2022

Thanks!

Q&A