JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

香港中文大學
The Chinese University of Hong Kong

ConQueR: Query Contrast Voxel-DETR for 3D Object Detection

CVPR 2023 Highlight

Poster: WED-AM-102

Project page: https://benjin.me/projects/2022_conquer/

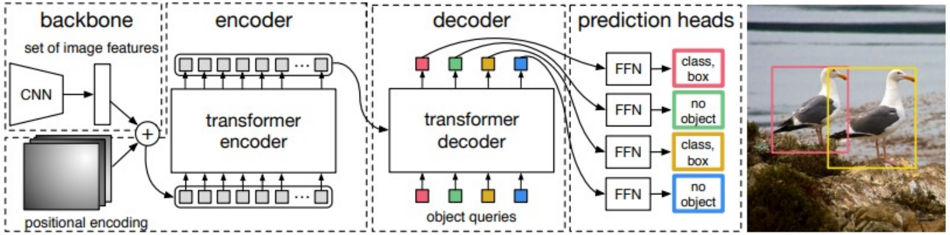
Benjin ZHU[1], *Zhe WANG*[2], *Shaoshuai SHI*[3], *Hang XU*[4], *Lanqing HONG*[4], *Hongsheng LI*[1]

[1]CUHK MMLab [2]SenseTime Research [3]Max Plank Institute for Informatics [4]Noah's Ark Lab

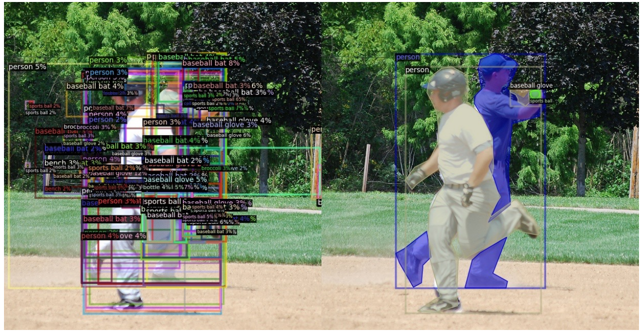
商汤 sensetime

Overview - Motivation

DETRs with direct **sparse** predictions have revolutionized object detection.



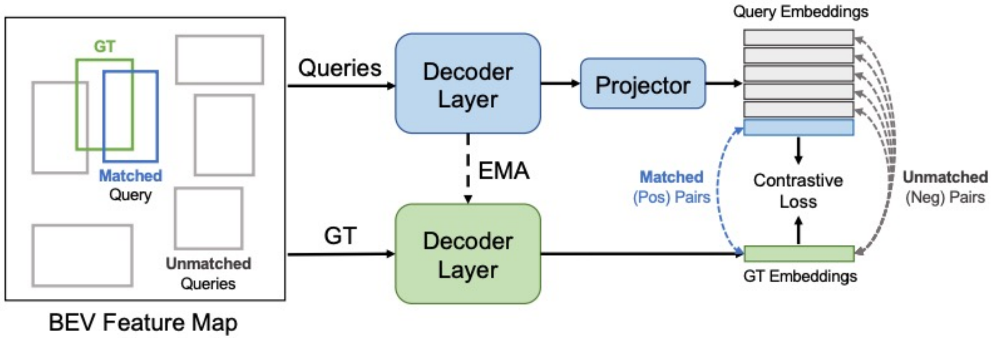
Fixed top-K outputting scheme causes **duplicated** predictions for both 2D and 3D DETRs.



100 predictions v.s. 4 objects (COCO)

Overview - Query Contrast

- Given the GT (green), Hungarian Matching gives its best matched (blue) and all other unmatched (gray) object queries. Query embeddings are projected by an extra MLP to align with GT embeddings. The contrastive loss is applied to all positive and negative GT-query pairs based on their feature similarities.

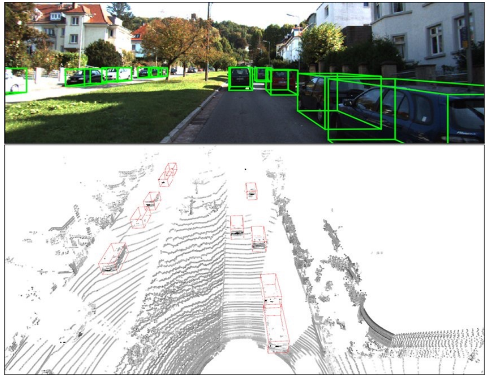


Objectives

- Identify all objects (category, 7 DoF box) w/o post-processing given point cloud inputs.
- Outperform dense detectors and achieve SOTA detection performance.

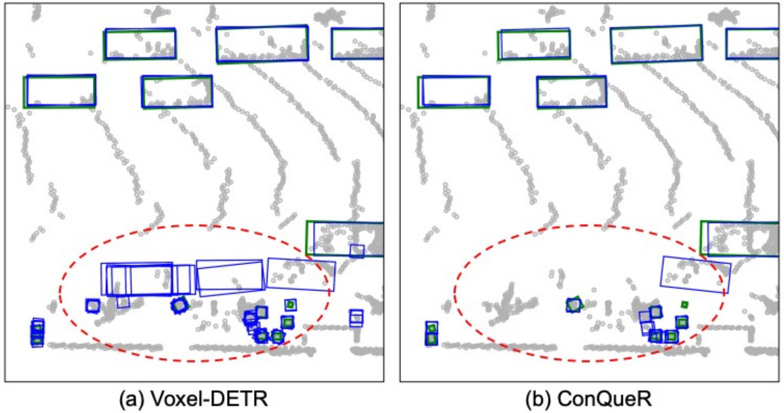
Previous works

1. Dense prediction with post-processing (e.g., NMS)
   - ✓ Strong performance.
   - ☹ Complex structures. Cannot be end-to-end optimized.
2. Direct sparse prediction
   - ✓ Clean pipeline. End-to-end optimizable.
   - ☹ Poor performance.



Motivation

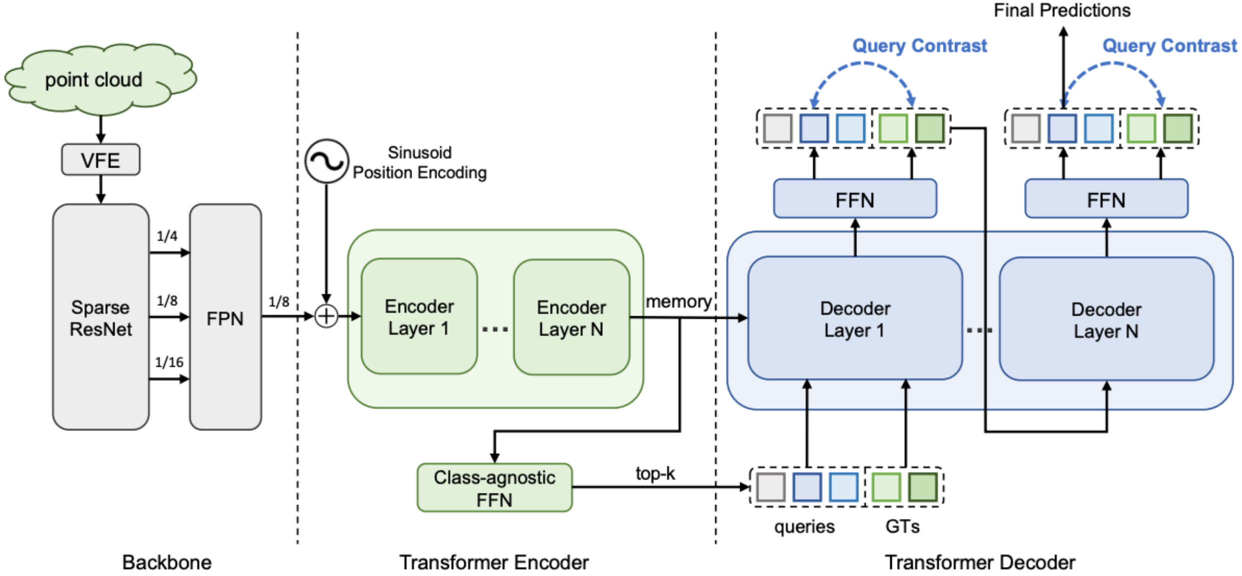
- DETRs usually adopt more queries than GTs (e.g., 300 queries v.s. ~40 objects in Waymo) in a scene, which inevitably incur many false positives during inference.
- Most false positives are highly overlapping in local regions, caused by the lack of explicit supervision in existing DETRs to discriminate and suppress locally similar queries.


(a) Voxel-DETR

(b) ConQueR

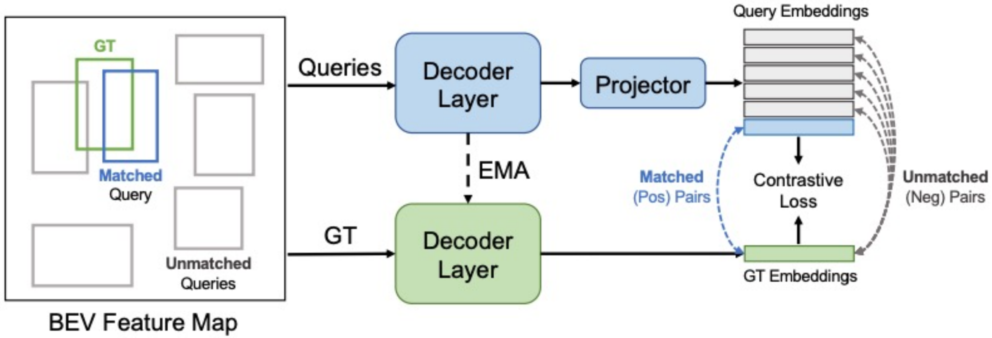
Strong Baseline: Voxel-DETR

- Clean architecture.
- Voxel-DETR with 6 epochs training already outperforms CenterPoint.



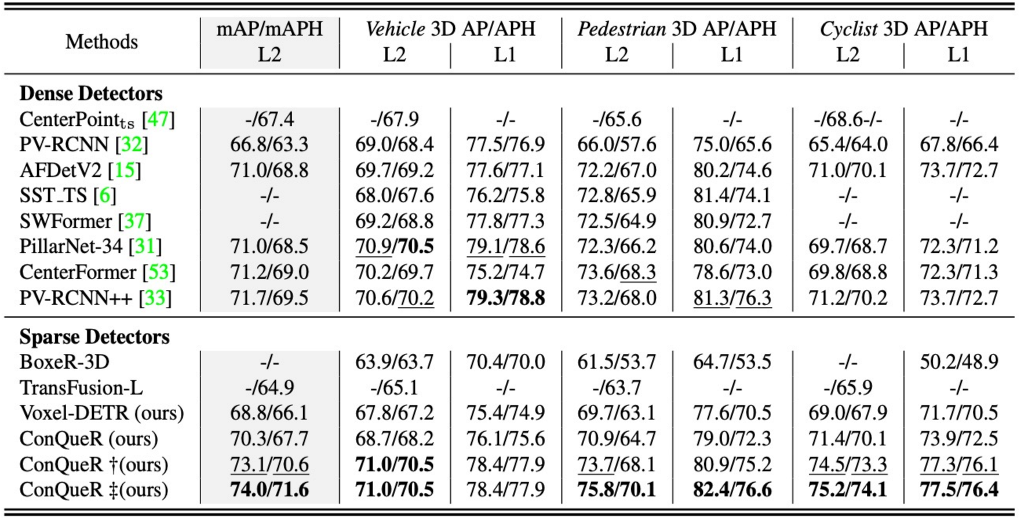
Query Contrast

- Construction of positive/negative GT-query pairs with Set Matching.
- Contrast positive GT-query pairs again negative ones with feature similarity-based contrastive loss.



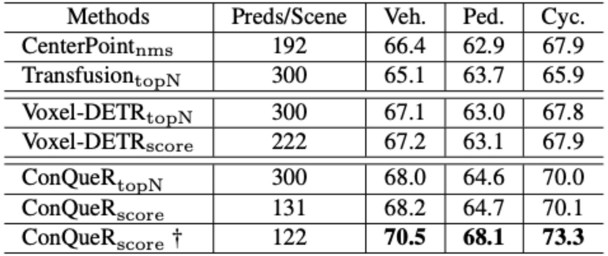
Experimental Results

- Query Contrast Voxel-DETR (ConQueR) significantly improves the detection performance and sparsity of final predictions, closes the gap between sparse and dense 3D detectors, and sets new records on the challenging WOD.



Methods	mAP/mAPH	*Vehicle* 3D AP/APH		*Pedestrian* 3D AP/APH		*Cyclist* 3D AP/APH	
	L2	L2	L1	L2	L1	L2	L1
Dense Detectors							
CenterPoint$_{\mathrm{ts}}$ [47]	-/67.4	-/67.9	-/-	-/65.6	-/-	-/68.6-/-	-/-
PV-RCNN [32]	66.8/63.3	69.0/68.4	77.5/76.9	66.0/57.6	75.0/65.6	65.4/64.0	67.8/66.4
AFDetV2 [15]	71.0/68.8	69.7/69.2	77.6/77.1	72.2/67.0	80.2/74.6	71.0/70.1	73.7/72.7
SST_TS [6]	-/-	68.0/67.6	76.2/75.8	72.8/65.9	81.4/74.1	-/-	-/-
SWFormer [37]	-/-	69.2/68.8	77.8/77.3	72.5/64.9	80.9/72.7	-/-	-/-
PillarNet-34 [31]	71.0/68.5	70.9/**70.5**	79.1/78.6	72.3/66.2	80.6/74.0	69.7/68.7	72.3/71.2
CenterFormer [53]	71.2/69.0	70.2/69.7	75.2/74.7	73.6/68.3	78.6/73.0	69.8/68.8	72.3/71.3
PV-RCNN++ [33]	71.7/69.5	70.6/70.2	**79.3/78.8**	73.2/68.0	81.3/76.3	71.2/70.2	73.7/72.7
Sparse Detectors							
BoxeR-3D	-/-	63.9/63.7	70.4/70.0	61.5/53.7	64.7/53.5	-/-	50.2/48.9
TransFusion-L	-/64.9	-/65.1	-/-	-/63.7	-/-	-/65.9	-/-
Voxel-DETR (ours)	68.8/66.1	67.8/67.2	75.4/74.9	69.7/63.1	77.6/70.5	69.0/67.9	71.7/70.5
ConQueR (ours)	70.3/67.7	68.7/68.2	76.1/75.6	70.9/64.7	79.0/72.3	71.4/70.1	73.9/72.5
ConQueR †(ours)	73.1/70.6	**71.0/70.5**	78.4/77.9	73.7/68.1	80.9/75.2	74.5/73.3	77.3/76.1
ConQueR ‡(ours)	**74.0/71.6**	**71.0/70.5**	78.4/77.9	**75.8/70.1**	**82.4/76.6**	**75.2/74.1**	**77.5/76.4**

Performance on the WOD validation split.



Methods	Preds/Scene	Veh.	Ped.	Cyc.
CenterPoint$_{\mathrm{nms}}$	192	66.4	62.9	67.9
Transfusion$_{\mathrm{topN}}$	300	65.1	63.7	65.9
Voxel-DETR$_{\mathrm{topN}}$	300	67.1	63.0	67.8
Voxel-DETR$_{\mathrm{score}}$	222	67.2	63.1	67.9
ConQueR$_{\mathrm{topN}}$	300	68.0	64.6	70.0
ConQueR$_{\mathrm{score}}$	131	68.2	64.7	70.1
ConQueR$_{\mathrm{score}}$ †	122	**70.5**	**68.1**	**73.3**

Sparsity of final predictions.