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

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

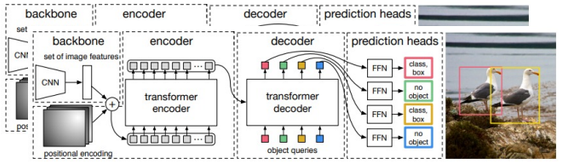
[1]CUHK Multimedia Lab [2]Max Plank Institute for Informatics [3]Huawei Noah's Ark Lab

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Limitations of DETRs

- ***Fixed top-N prediction*** in DETR-based detectors cause highly overlapping false positives in local regions.

- ***Inter-query relations*** are not considered in the Set-Matching loss of DETRs. → Limited capabilities in discriminating local similar queries.



Objectives

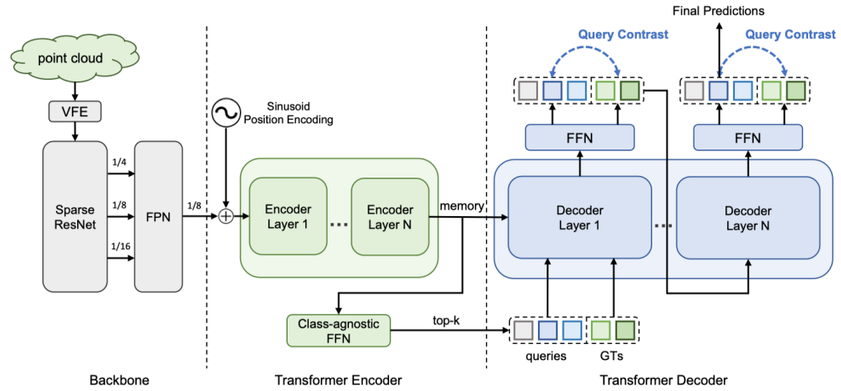
Status Quo of 3D Object Detection

- Dense prediction with post-processing (e.g., NMS)
  - ✓ Strong performance.
  - ☹ Complex structures. Not e2e optimizable.
- Direct sparse prediction (e.g., DETRs)
  - ✓ Clean pipeline. End-to-end optimizable.
  - ☹ Poor performance.

ConQueR achieves

- Direct **Sparse** Prediction with **Strong** Performance.
- **Less** overlapped false positives.
- **Dynamic** #predictions according to scene.

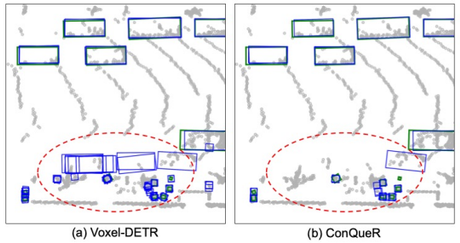
Voxel-DETR – A Simple and Strong Baseline



- **Simple and clean pipeline, end-to-end optimizable.**
- **Strong performance**: Surpasses all previous sparse 3D detectors by a large margin.
- **Fast convergence**: comparable with CenterPoint with 1/6 training time (epochs).

Observations

- Most false positives are highly overlapping in local regions, caused by the lack of explicit supervision to discriminate locally similar queries.

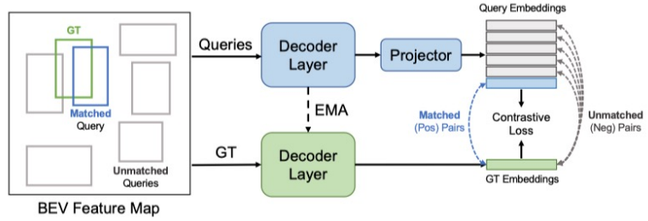


(a) Voxel-DETR (b) ConQueR

- In the Hungarian Matching Loss of existing 3D DETRs, the best matched query are supervised **without considering its relative ranking to its surrounding unmatched queries.**

Query Contrast

Explicitly **enhance** each GT's best matched query, and **suppress** predictions from all other unmatched ones with contrastive loss **simultaneously**.



Experimental Results

Performance

Methods	mAP/mAPH L2	*Vehicle* 3D AP/APH L2	*Vehicle* 3D AP/APH L1	*Pedestrian* 3D AP/APH L2	*Pedestrian* 3D AP/APH L1	*Cyclist* 3D AP/APH L2	*Cyclist* 3D AP/APH L1
Dense Detectors							
CenterPoint$_{ts}$ [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**

Sparsity

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

contact code