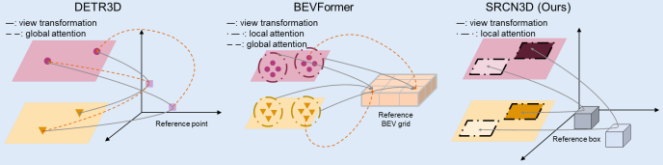
SRCN3D: Sparse R-CNN 3D for Compact Convolutional Multi-View 3D Object Detection and Tracking

Yining Shi, Jingyan Shen, Yifan Sun, Yunlong Wang, Jiaxin Li, Shiqi Sun, Kun Jiang*, Diange Yang*
Tsinghua University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction



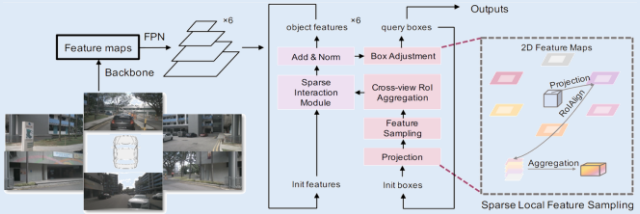
Motivation

Vision-based multi-view 3D (MV3D) detection and tracking becomes a new trend in autonomous perception. A sparse paradigm is likely to benefit MV3D efficiency.

Technical Insight

- We develop a transformer-less network-based MV3D method to demonstrate that CNNs are equally capable of performing view transformations in MV3D.
- We develop a box-wise two-stage cascade detector that performs sparse feature sampling compared to point-wise sampling or dense query-based detectors.

Overview of SRCN3D



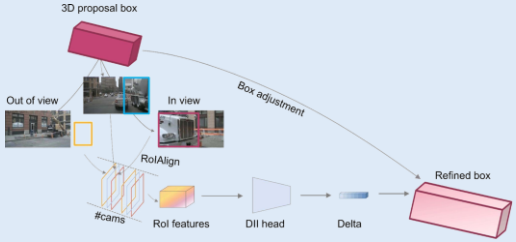
Key points

- SRCN3D predicts **3D bounding boxes directly in 3D world space** without depth supervision and post-processing like non-maximum suppression.
- As **a sparse paradigm**, each query box serves as a filter to focus on a sparse local region of 2D feature maps.
- **A fully-convolutional pipeline** is designed without mask operations, positional embeddings and attention weights in typical vision transformers.

SRCN3D Head:

Sparse feature sampling module

- 3D query boxes: a fixed number of boxes parameterized to the same dimension as 3D bounding box.
- Query features: sets of high-dimensional latent vectors, strictly corresponding to 3D query boxes.
- Cross-view fusion: aggregation of projected RoI features. RoI features maintain a fixed expression, no matter how many cameras capture one query box.



Sparse interaction head

In sparse interaction head, RoI features extracted from query boxes are passed through two 1x1 convolutional layers for interaction, followed by a Feed-Forward Network (FFN) block with layer normalization and a linear projection block to output classification and regression predictions. Corresponding parameters are generated from query features via linear transformation to achieve local interaction.

Box adjustment

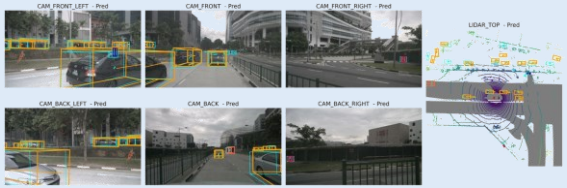
Query boxes are updated in each stage through box adjustment to refine the predictions.

Code is available at: https://github.com/synsin0/SRCN3D.

Experiments and Results:

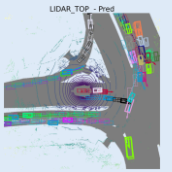
Result and visualization on nuScenes detection benchmark:

Method	Size	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS↑
CenterNet [8]	-	DLA	0.328	0.306	0.716	0.264	0.609	1.426	0.658	-
FCOS3D †# [27]	1600×900	Res-101	0.415	0.343	0.725	**0.263**	0.422	1.292	**0.153**	2.0
DETR3D ¶ [28]	1600×900	Res-101	0.425	0.346	0.773	0.268	0.383	0.842	0.216	2.7
BEVDet § [14]	1056×384	Res-101	0.396	0.330	**0.702**	0.272	0.534	0.932	0.251	**16.7**
BEVFormer-S ¶ [16]	1600×900	Res-101	**0.448**	**0.375**	0.725	0.272	0.391	0.802	0.200	2.1
PETR ¶‖ [19]	1600×900	Res-101	0.442	0.370	0.711	0.267	0.383	0.865	0.201	2.5
SRCN3D (Ours)‖	1600×900	Res-101	0.428	0.337	0.779	0.287	**0.367**	**0.781**	0.188	3.2
SRCN3D (Ours)‖	1600×900	V2-99	0.475	0.396	0.737	0.294	0.278	0.728	0.197	2.5



Result and visualization on nuScenes tracking benchmark:

Method	Modality	Split	AMOTA↑	AMOTP↓	RECALL↑
QD3DT [12]	C	val	0.242	1.518	0.399
MUTR3D [32]	C	val	0.294	1.498	0.427
ViP3D [10]	C	val	0.216	1.616	0.358
UniAD [13]	C	val	0.359	1.320	0.467
SRCN3D (Ours)	C	val	**0.439**	**1.280**	**0.545**
CenterTrack-Open [33]	L + C	test	0.108	**0.989**	0.412
QD-3DT [12]	C	test	0.217	1.550	0.375
PolarDETR [6]	C	test	0.273	1.185	0.404
DEFT [1]	C	test	0.177	1.564	0.338
MUTR3D [32]	C	test	0.270	1.494	0.411
SRCN3D (Ours)	C	test	**0.398**	1.317	**0.538**



Contributions

- **The first transformer-less two-stage MV3D approach with box-wise sampling**, leading to a more straightforward, lightweight and faster detection pipeline.
- **A novel sparse cross-attention module to refine 3D queries from 2D feature maps** with a lower computation cost is achieved.
- Extensive experiments on nuScenes dataset demonstrate the effectiveness of SRCN3D for 3D object detection and tracking.