An End-to-End Transformer Model for 3D Object Detection

Ananya, Andrew, Kaan, Sizhe

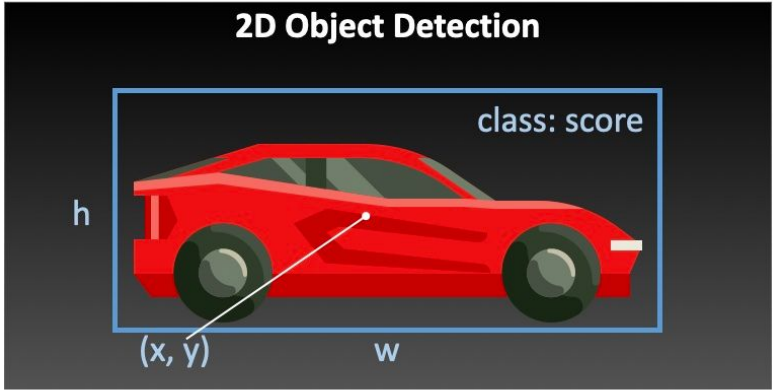
Motivation

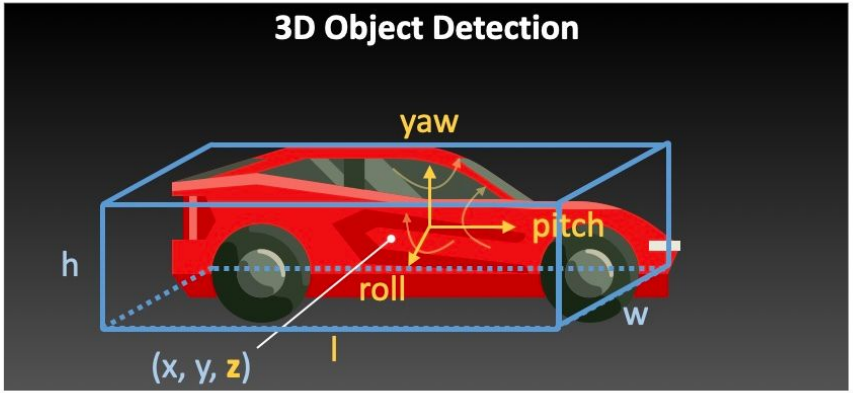
3D Object Detection

- 3D object detection can produce 3D bounding boxes
- Important for medical imaging, autonomous vehicles, augmented reality

- Existing 3D detection models are carefully tuned and include inductive biases

- 3DETR presents a simple and accurate model for 3D object detection

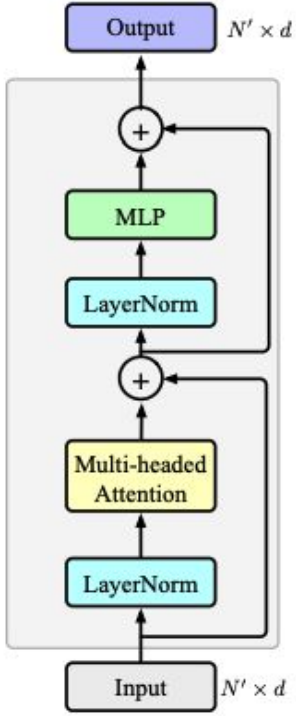




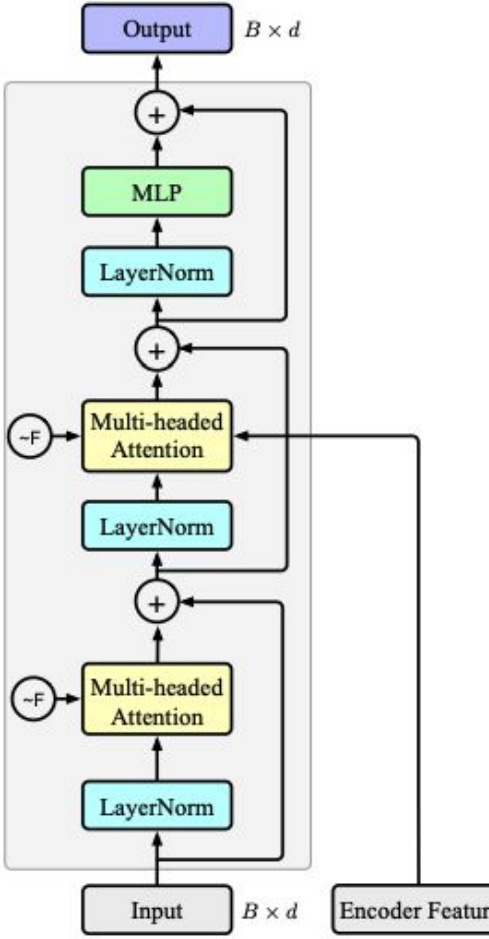
Method

3DETR Architecture

- Input: 3D point cloud
- Output: 3D bounding boxes with labels
- Out-of-the-box Transformer design



One Layer of Transformer Encoder



One Layer of Transformer Decoder

Implementation and Training

Framework	PyTorch
Optimizer	AdamW
Augmentation	RandomCuboid
Training GPU	8*V100
Epochs	1080

Empirical Results

Experimental Setup

- ScanNetV2
  - 1.2K point cloud samples, 18 object categories
- SUN RGB-D-v1
  - 5K point cloud samples, 37 object categories

- Test compared to BoxNet and VoteNet
  - Foundational for recent detection models
- Authors re-implemented these models for fair comparison
  - Led to 2 to 4% increase over original paper

Results - SOTA Comparison

- 3DETR-m (3DETR with **masked** self-attn.) outperforms both BoxNet and VoteNet.
- Achieve competitive results comparing to H3DNet.

Method	ScanNetV2		SUN RGB-D	
	AP_{25}	AP_{50}	AP_{25}	AP_{50}
BoxNet† [42]	49.0	21.1	52.4	25.1
3DETR	62.7	37.5	58.0	30.3
VoteNet† [42]	60.4	37.5	58.3	33.4
3DETR-m	65.0	47.0	59.1	32.7
H3DNet [89]	67.2	48.1	60.1	39.0

* H3DNet: A 3D-specific architecture based on VoteNet

Results - Encoder Comparison

- 3DETR's encoder is more effective than PointNet++

Method	Encoder	Decoder	Loss	ScanNetV2		SUN RGB-D	
				AP_{25}	AP_{50}	AP_{25}	AP_{50}
3DETR	Tx.	Tx.	Set	62.7	37.5	58.0	30.3
	PN++	Tx.	Set	61.4	34.7	56.8	26.9

PN++: PointNet++ [45], Tx.: Transformer, Set loss § 3.4

Results - Decoder & Loss Comparison

- 3DETR's Decoder is more effective than that used by VoteNet and BoxNet
- 3DETR's set loss also made a significant difference.

#	Method	Encoder	Decoder	Loss	ScanNetV2		SUN RGB-D	
					AP_{25}	AP_{50}	AP_{25}	AP_{50}
Comparing different decoders								
1	3DETR	Tx.	Tx.	Set	62.7	37.5	58.0	30.3
2		Tx.	Box	Box	31.0	10.2	36.4	14.4
3		Tx.	Vote	Vote	46.1	23.4	47.5	24.9
Comparing different losses								
4		Tx.	Tx.	Box	49.6	20.5	49.5	21.1
5		Tx.	Tx.	Vote	54.0	31.9	53.4	28.3

Tx.: Transformer, Vote/Box loss [42], Set loss § 3.4

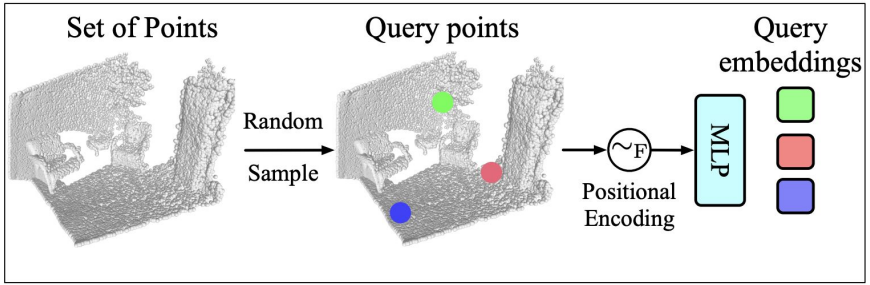
Results - Shape Classification

- Shows that 3DETR can be flexibly adapted to various tasks and achieve competitive results.

Method	input	mAcc	OA
PointNet++ [45]	point	–	91.9
SpecGCN [71]	point	–	92.1
DGCNN [77]	point	90.2	92.2
PointWeb [90]	point	89.4	92.3
SpiderCNN [80]	point	–	92.4
PointConv [78]	point	–	92.5
KPConv [67]	point	–	92.9
InterpCNN [34]	point	–	93.0
3DETR encoder (Ours)	point	89.1	92.1
3DETR-m encoder (Ours)	point	89.9	91.9

Results - Positional Embedding & Query

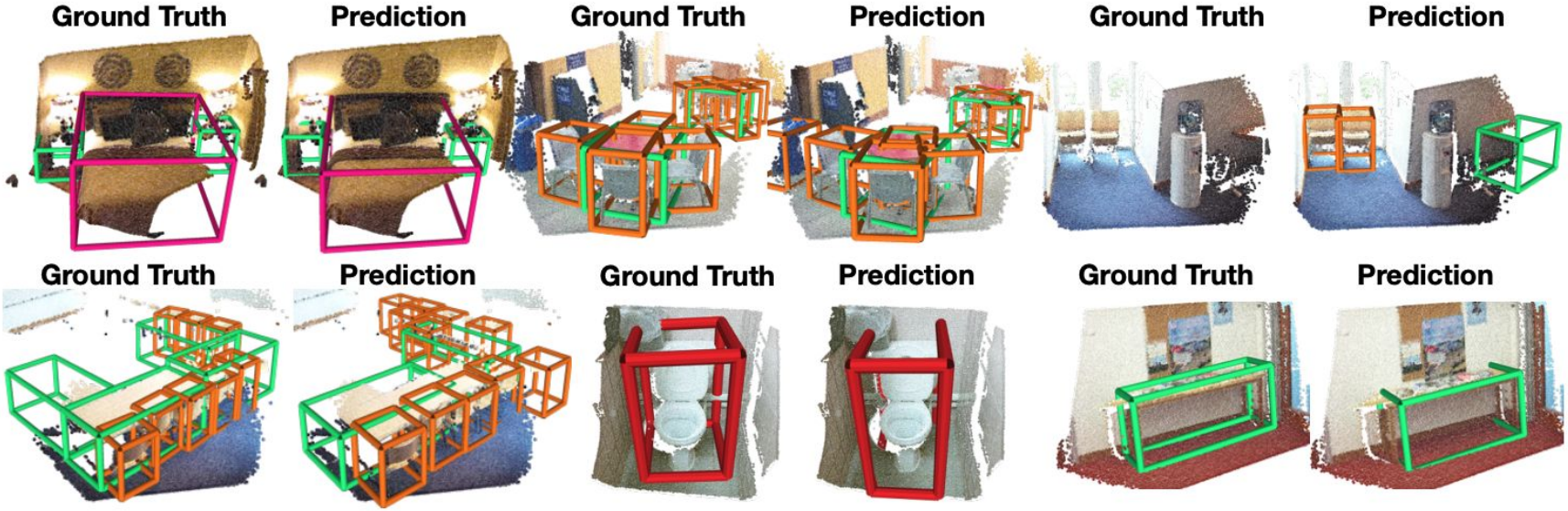
- Clear benefit for non-parametric query embedding and Fourier positional encoding



#	Method	Positional Embedding		Query Type	ScanNetV2	
		Encoder	Decoder		AP_{25}	AP_{50}
1	3DETR	-	Fourier	np + Fourier	62.7	37.5
2		Fourier	Fourier	np + Fourier	61.8	37.0
3		Sine	Sine	np + Sine	55.8	30.9
4		-	-	np + Sine	31.3	10.8
5	DETR [4]†	Sine	Sine	parametric [4]	15.4	5.3

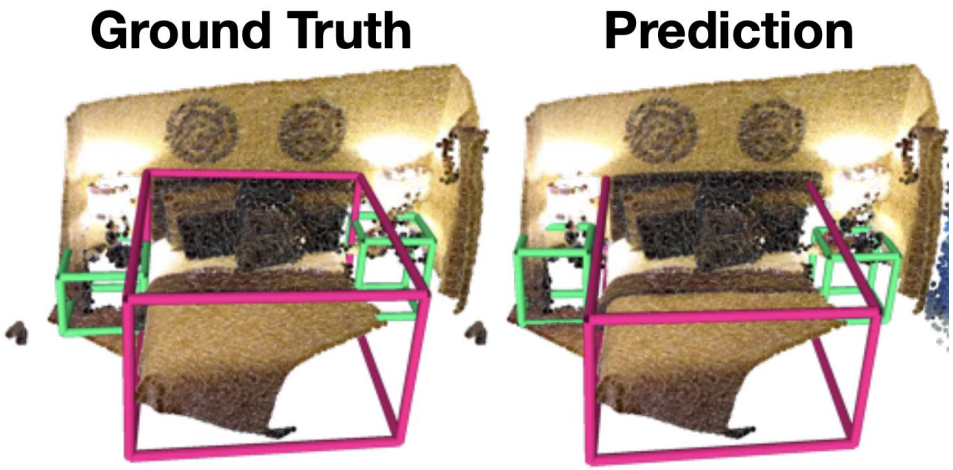
np: non-parametric query (§ 3.2)

Handpicked Examples

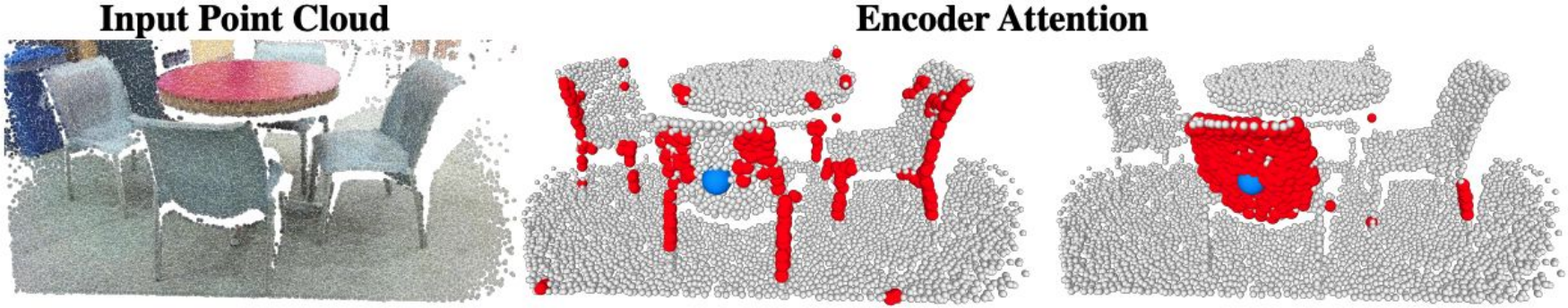


Key Examples: Occluded Objects

- Corner of bed not visible in scan
- Important for single-view applications



Visualizing Attention



- Reference point in blue
- Points with highest attention in red
- Each head focuses on different geometric parts

Detailed Comparison to State-of-the-Art

Method	Arch.	ScanNetV2		SUN RGB-D	
		AP_{25}	AP_{50}	AP_{25}	AP_{50}
BoxNet† [42]	BoxNet	49.0	21.1	52.4	25.1
3DETR	Tx.	62.7	37.5	56.8	30.1
VoteNet† [42]	VoteNet	60.4	37.5	58.3	33.4
3DETR-m	Tx.	65.0	47.0	59.0	32.7
H3DNet [89]	VoteNet + 3D primitives	67.2	48.1	60.1	39.0
HGNet [5]	VoteNet + GraphConv	61.3	34.4	61.6	34.4
3D-MPA [11]	VoteNet + GraphConv	64.2	49.2	-	-

Table 11: Detailed state-of-the-art comparison on 3D detection.