End-to-End 3D Dense Captioning with Vote2Cap-DETR

Poster Session: WED-AM-275

Sijin Chen[1] Hongyuan Zhu[2] Xin Chen[3] Yinjie Lei[4] Gang YU[3] Tao Chen[1†]

[1] Fudan University [2] Institute for Infocomm Research (I2R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3] Tencent PCG [4] Sichuan University

[†] Corresponding author.

Paper Code Fudan EDL Lab

Contact Us: eetchen@fudan.edu.cn

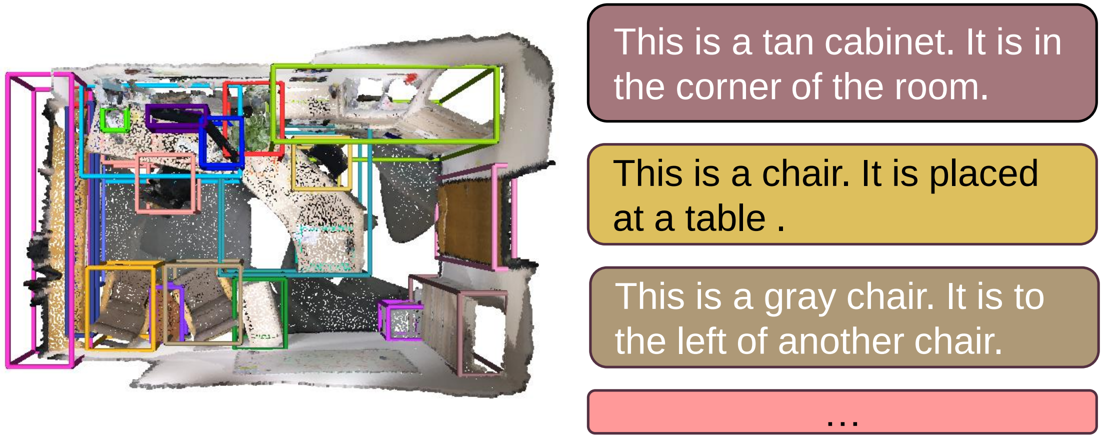
3D Dense Captioning

3D Dense Captioning requires:

1. Accurate localization of all objects of interests in a 3D scene;
2. Informative and object-centric descriptions for each instance.

Sparse and Cluttered 3D Scene

3D Dense Captioning



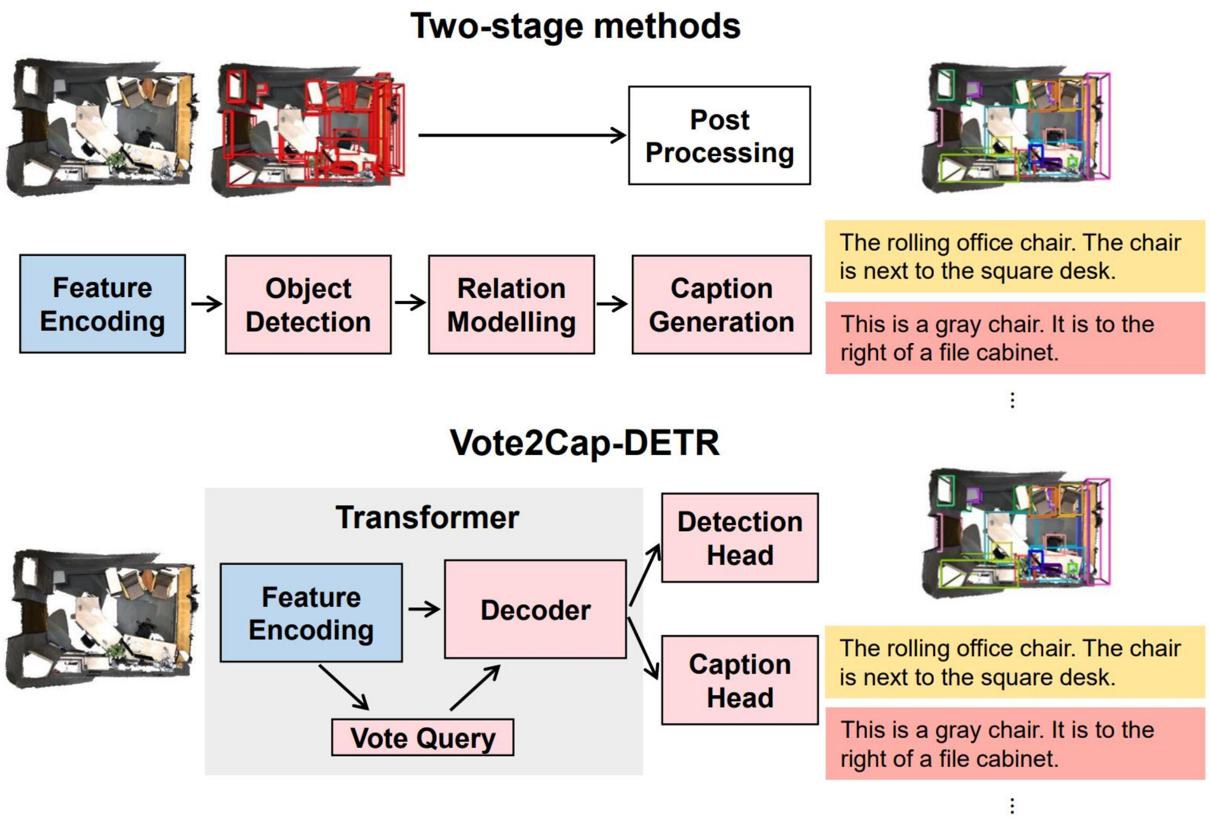
Motivation

Previous "detect-then-describe" pipeline:

1. the whole model relies on the detector's output;
2. huge amount of hand-crafted components.

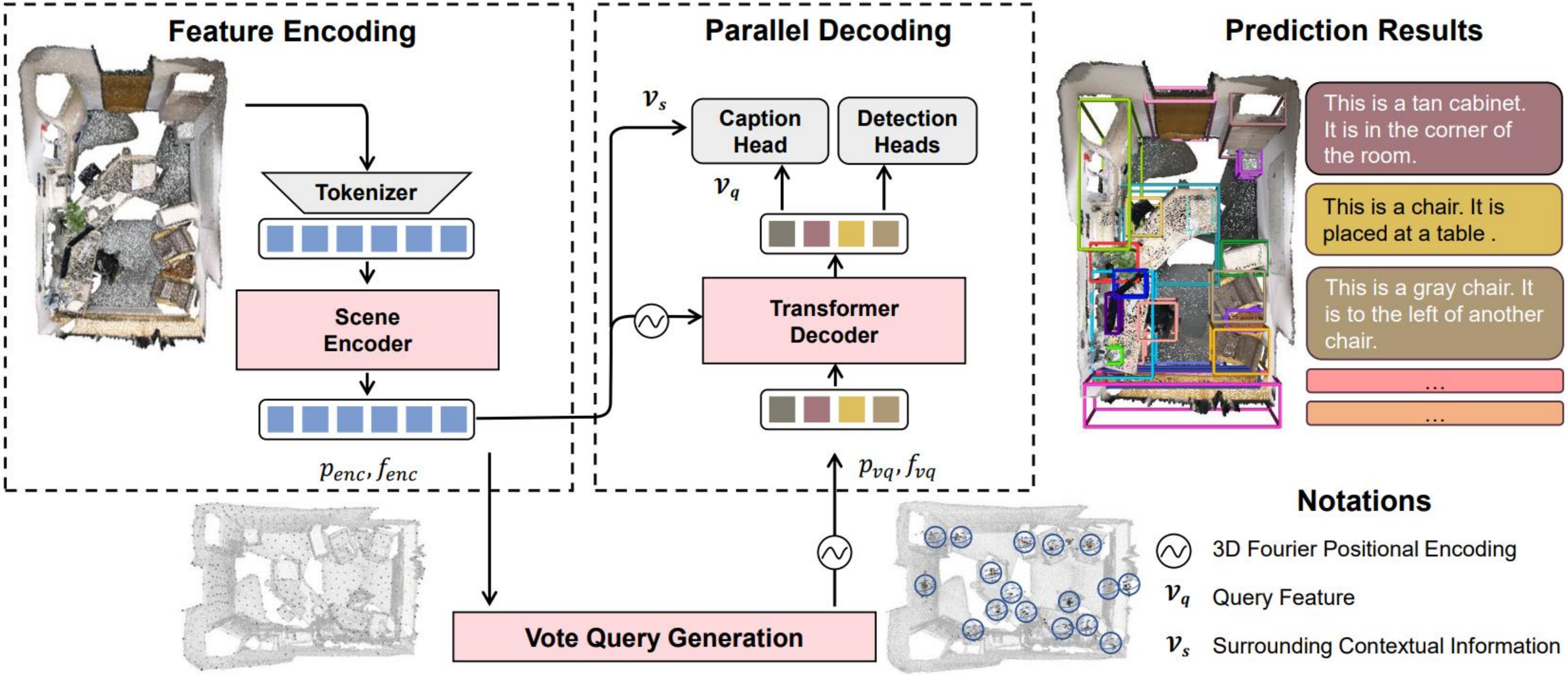
A one-stage system (Ours):

1. single-stage transformer-based architecture;
2. parallel decoding queries to boxes and captions.



Methods

- Vote2Cap-DETR Overview



Methods

➢ Vote Query

We reformulate object queries as $(xyz, feat)$.

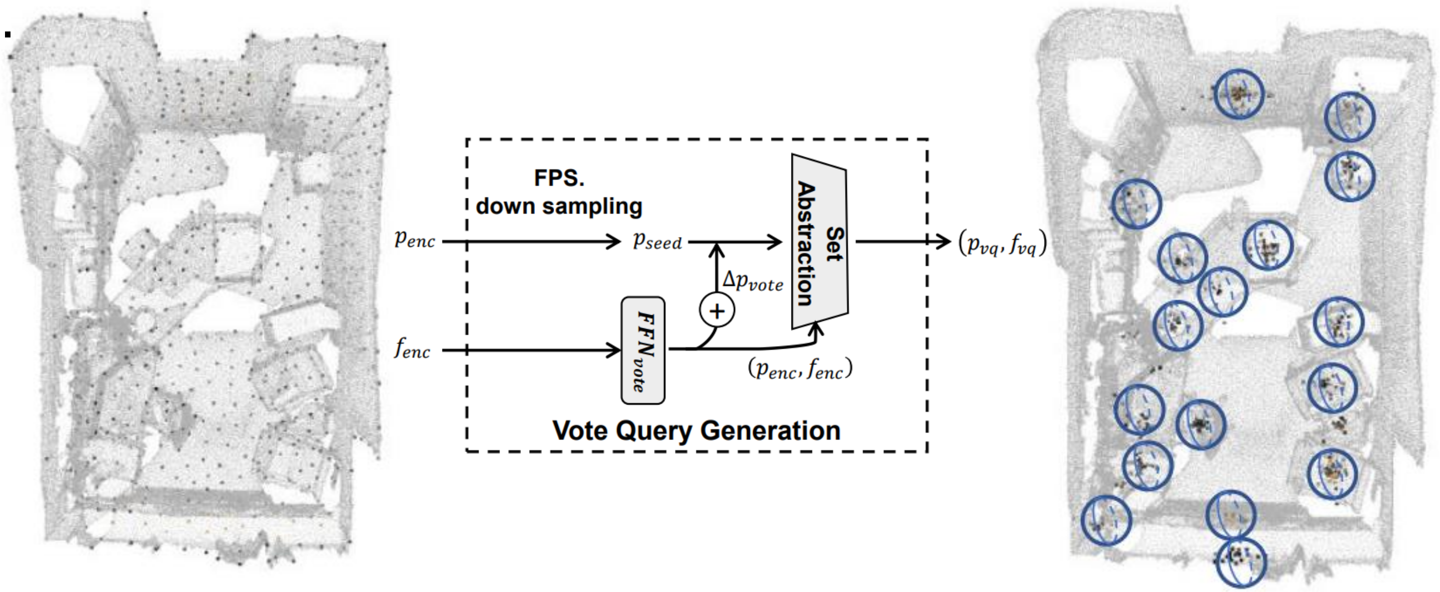
$$f^{i}_{query} = Layer_{i-1}\left(f^{i-1}_{query} + FFN\left(PE\left(p_{vq}\right)\right)\right)$$

Vote queries learn to:

1. shift seed points to probable locations of objects;

$$p_{vote} = p_{enc} + \Delta p_{vote} = p_{enc} + FFN_{vote}\left(f_{enc}\right)$$

2. aggregate feature from the local context.

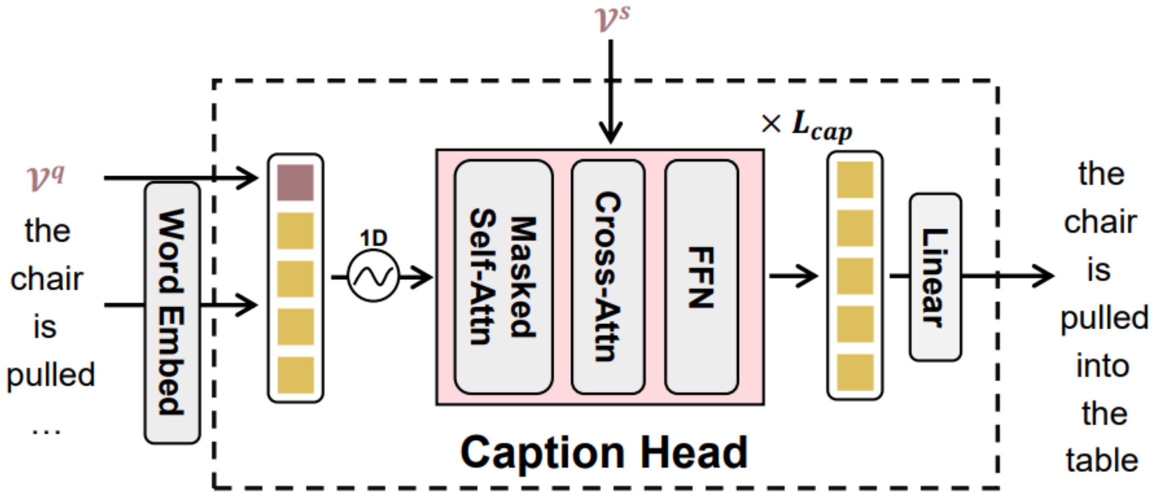


Methods

➢ Dual Clued Captioner

To generates informative, and object-centric captions for objects, the captioner receives two streams of visual clues:

1. the object vote feature $\mathcal{V}^q$ to identify the object;
2. looking into the local context $\mathcal{V}^s$ surrounding the query $\mathcal{V}^q$.



Quantitative Results

- ScanRefer validation set

Method	$\mathcal{L}_{des}$	w/o additional 2D input								w/ additional 2D input							
		IoU = 0.25				IoU = 0.50				IoU = 0.25				IoU = 0.50			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap [13]	MLE	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
MORE [20]	MLE	58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3d [39]	MLE	58.06	35.30	26.16	55.03	42.76	25.38	22.84	45.66	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG [4]	MLE	60.86	**39.67**	27.45	59.02	47.68	31.53	24.28	51.80	64.70	**40.17**	27.66	**59.23**	49.48	31.03	24.22	50.80
D3Net [7]	MLE	-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
Ours	MLE	**71.45**	39.34	**28.25**	**59.33**	**61.81**	**34.46**	**26.22**	**54.40**	**72.79**	39.17	**28.06**	**59.23**	**59.32**	**32.42**	**25.28**	**52.53**
χ-Trans2Cap [43]	SCST	58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
Scan2Cap [13]	SCST	-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net [7]	SCST	-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	**25.72**	**53.90**
Ours	SCST	**84.15**	**42.51**	**28.47**	**59.26**	**73.77**	**38.21**	**26.64**	**54.71**	**86.28**	**42.64**	**28.27**	**59.07**	**70.63**	**35.69**	25.51	52.28

- Nr3D validation set

Method	$\mathcal{L}_{des}$	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap [13]	MLE	27.47	17.24	21.80	49.06
SpaCap3d [39]	MLE	33.71	19.92	22.61	50.50
D3Net [7]	MLE	33.85	20.70	23.13	53.38
3DJCG [4]	MLE	38.06	22.82	23.77	52.99
Ours	MLE	**43.84**	**26.68**	**25.41**	**54.43**
χ-Tran2Cap [43]	SCST	33.62	19.29	22.27	50.00
D3Net [7]	SCST	38.42	22.22	24.74	54.37
Ours	SCST	**45.53**	**26.88**	**25.43**	**54.76**

Quantitative Results

Scan2Cap Benchmark

This table lists the benchmark results for the Scan2Cap Dense Captioning Benchmark scenario.

		Captioning F1-Score				Dense Captioning	Object Detection
Method	Info	CIDEr@0.5IoU	BLEU-4@0.5IoU	Rouge-L@0.5IoU	METEOR@0.5IoU	DCmAP	mAP@0.5
vote2cap-detr		**0.3128** 1	**0.1778** 1	**0.2842** 1	**0.1316** 1	**0.1825** 1	**0.4454** 1
CFM		0.2360 2	0.1417 2	0.2253 2	0.1034 2	0.1379 5	0.3008 5
CM3D-Trans+		0.2348 3	0.1383 3	0.2250 4	0.1030 3	0.1398 4	0.2966 7
Yufeng Zhong, Long Xu, Jiebo Luo, Lin Ma: Contextual Modeling for 3D Dense Captioning on Point Clouds.							
Forest-xyz		0.2266 4	0.1363 4	0.2250 3	0.1027 4	0.1161 10	0.2825 10
D3Net - Speaker	P	0.2088 5	0.1335 6	0.2237 5	0.1022 5	0.1481 3	0.4198 2
Dave Zhenyu Chen, Qirui Wu, Matthias Niessner, Angel X. Chang: D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. 17th European Conference on Computer Vision (ECCV), 2022							
3DJCG(Captioning)	P	0.1918 6	0.1350 5	0.2207 6	0.1013 6	0.1506 2	0.3867 3
Daigang Cai, Lichen Zhao, Jing Zhang†, Lu Sheng, Dong Xu: 3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds. CVPR2022 Oral							
REMAN		0.1662 7	0.1070 7	0.1790 7	0.0815 7	0.1235 8	0.2927 9

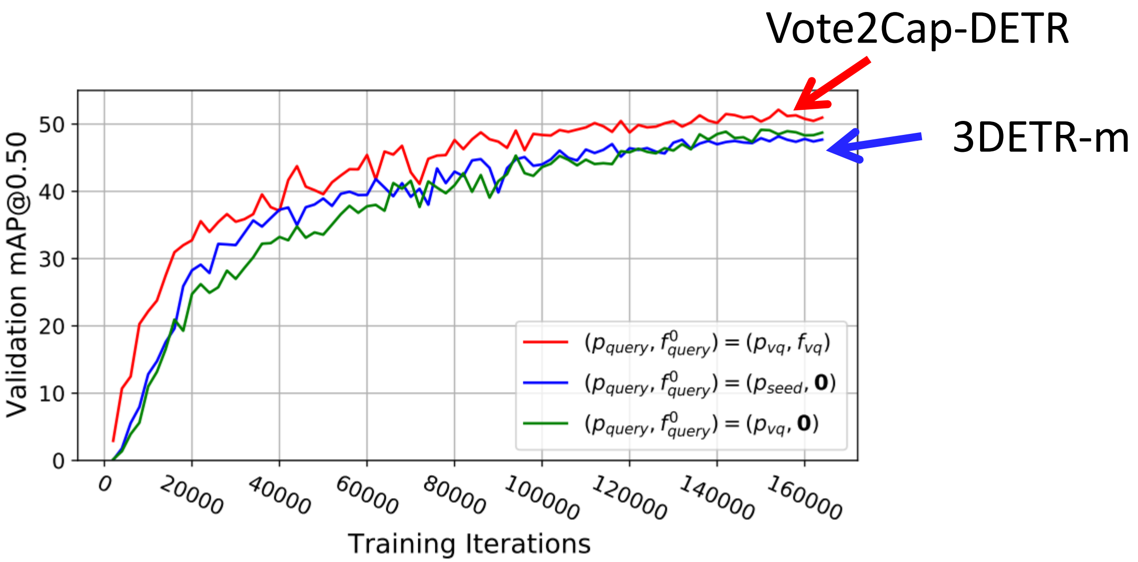
* Ranked 1st on the ScanRefer online test benchmark, https://kaldir.vc.in.tum.de/scanrefer_benchmark/benchmark_captioning

Study on Components

➢ Does the vote query improve 3DETR?

Comparison to other 3DETR attempts. We compare detection performance of different methods that improve 3DETR in the 20k, 40k, 80k, 160k -th iteration.

Model	Modification	(20k)AP@0.5↑	(40k)AP@0.5↑	(80k)AP@0.5↑	(160k)AP@0.5↑
3DETR-m	-	28.26	37.27	43.41	48.18
3DETR-m	hybrid	**35.10**	**42.72**	45.83	47.50
3DETR-m	anchor	22.94	28.85	35.44	40.06
Vote2Cap-DETR	-	32.70	40.90	**47.62**	**52.49**



➢ Does 3D context feature help captioning?

Different keys for caption generation. Introducing local contextual information leads to more informative and object-centric captions.

key	IoU=0.25				IoU=0.5			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
-	68.62	38.61	27.67	58.47	60.15	34.02	25.80	53.82
global	70.05	39.23	27.84	58.44	61.20	34.66	25.93	53.79
local	**70.42**	**39.98**	**27.99**	**58.89**	**61.39**	**35.24**	**26.02**	**54.12**

Qualitative Results

scene0011_00

3DJCG: This is a rectangular whiteboard. It is on the wall.

SpaCap3D: The whiteboard is affixed to the wall. It is to the right of the window.

Ours: The tv is on the wall. It is to the right of the table.

GT: This is a big black tv. It is above a thin table.

scene0015_00

3DJCG: This is a brown table. It is in the middle of the room.

SpaCap3D: This is a wooden table. It is in the center of the room.

Ours: This is a wooden table. It is in the corner of the room.

GT: This is a small table with a wood look. It is the table closest to the front of the room in the upper left corner.

scene0025_00

3DJCG: The is a small brown cabinet. It is to the right of the desk.

SpaCap3D: The cabinet is below the desk. It is to the left of the chair.

Ours: This is a white cabinet. It is to the right of the table.

GT: A white cabinet is sitting on the floor next to the wall. It is to the left of the couch.

scene0050_00

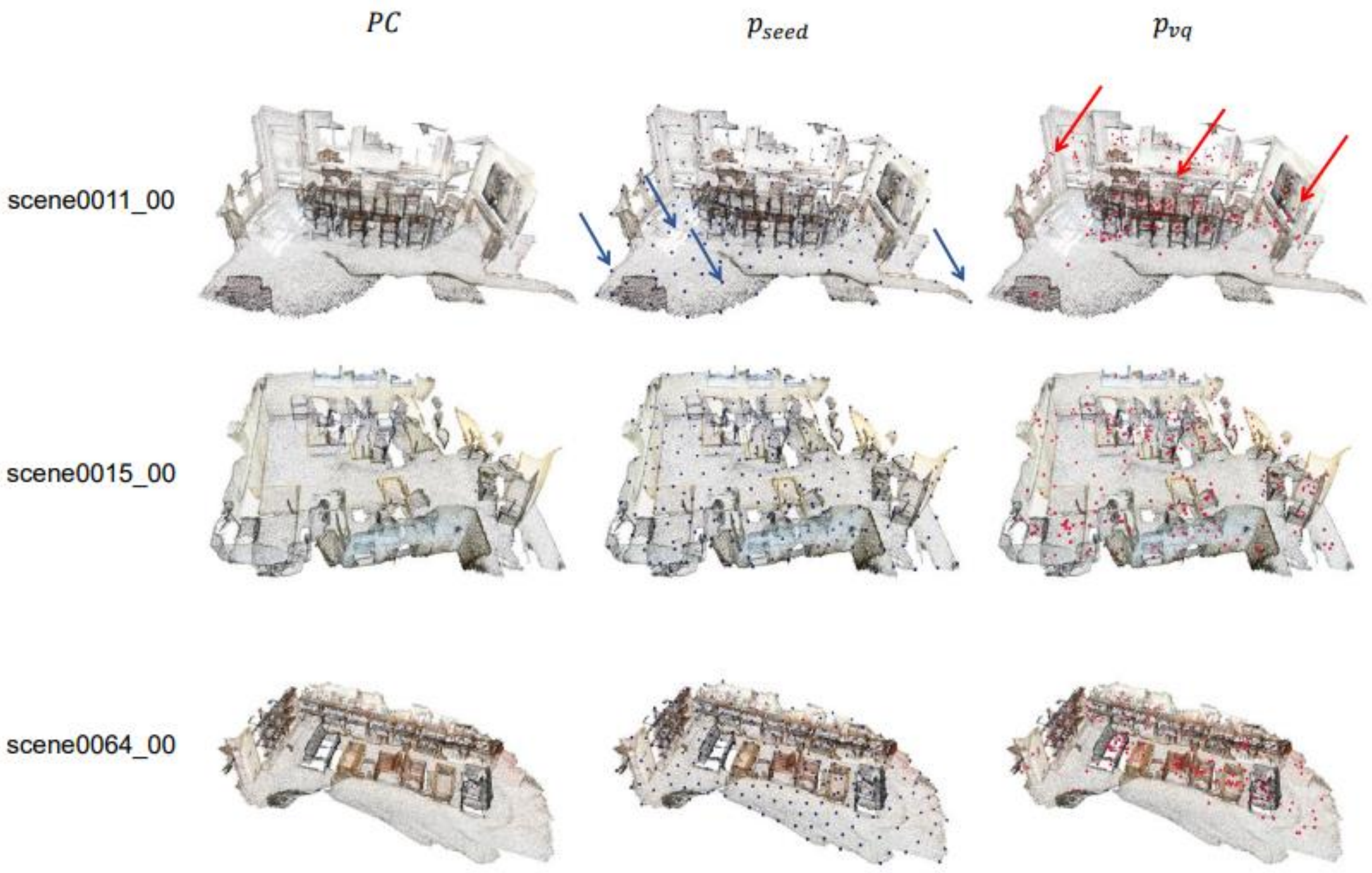
3DJCG: This is a brown table. It is in front of the couch.

SpaCap3D: This is a wooden coffee table. It is in front of the couch.

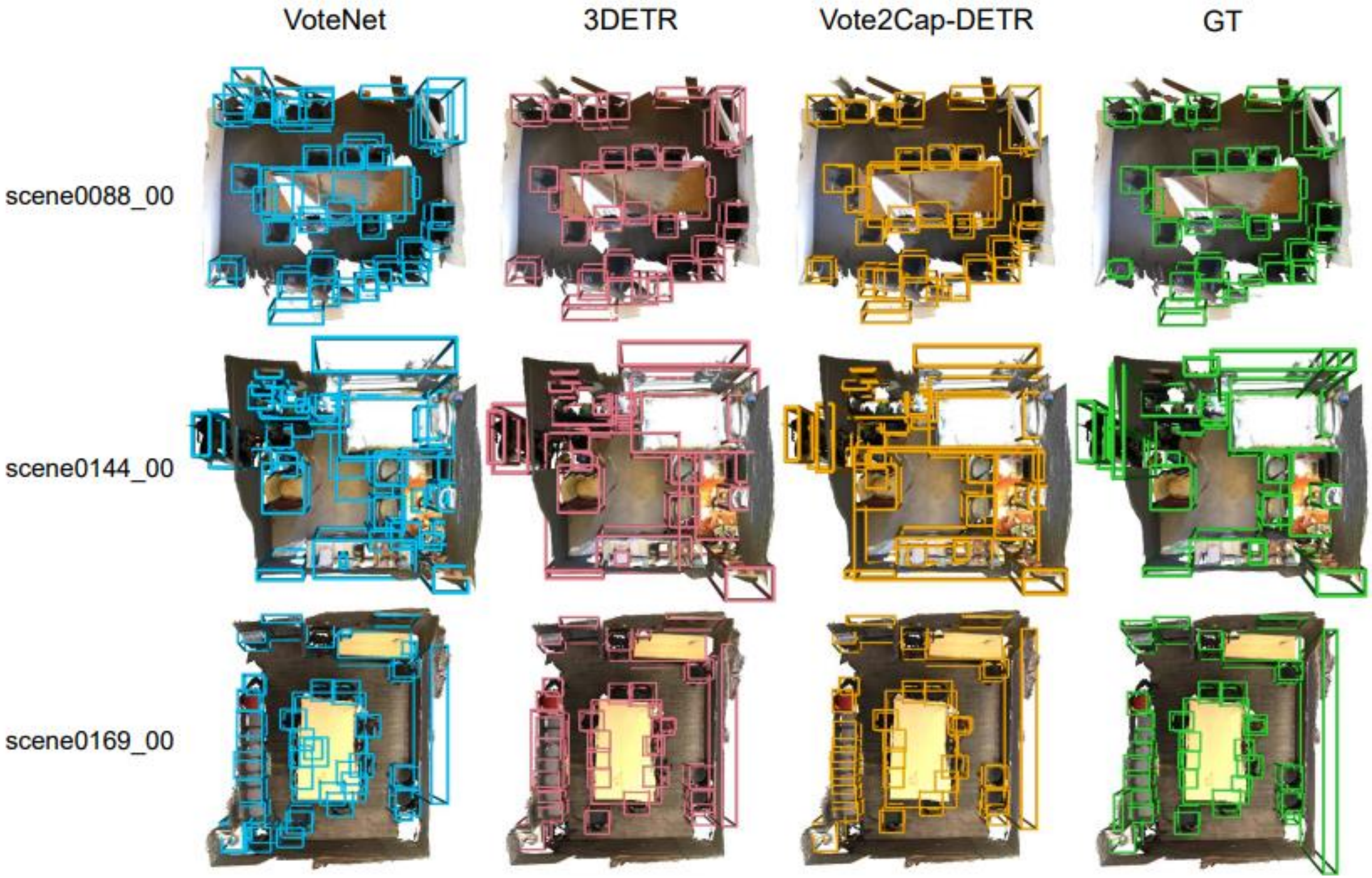
Ours: This is a brown ottoman. It is to the right of the chair.

GT: This is a brown ottoman. It is in front of a couch.

Visualization: Vote Queries



Visualization: Detection Results



Conclusions

We introduce a novel one-stage method to 3D dense captioning:

1. By introducing spatial bias and content-aware features, **vote queries** boost both convergence and detection performance.
2. The novel lightweight caption head looks into both query feature and **local contexts** for informative caption generation.

Thanks!

Paper

Code

Fudan EDL Lab