Bi-directional Contextual Attention for 3D Dense Captioning

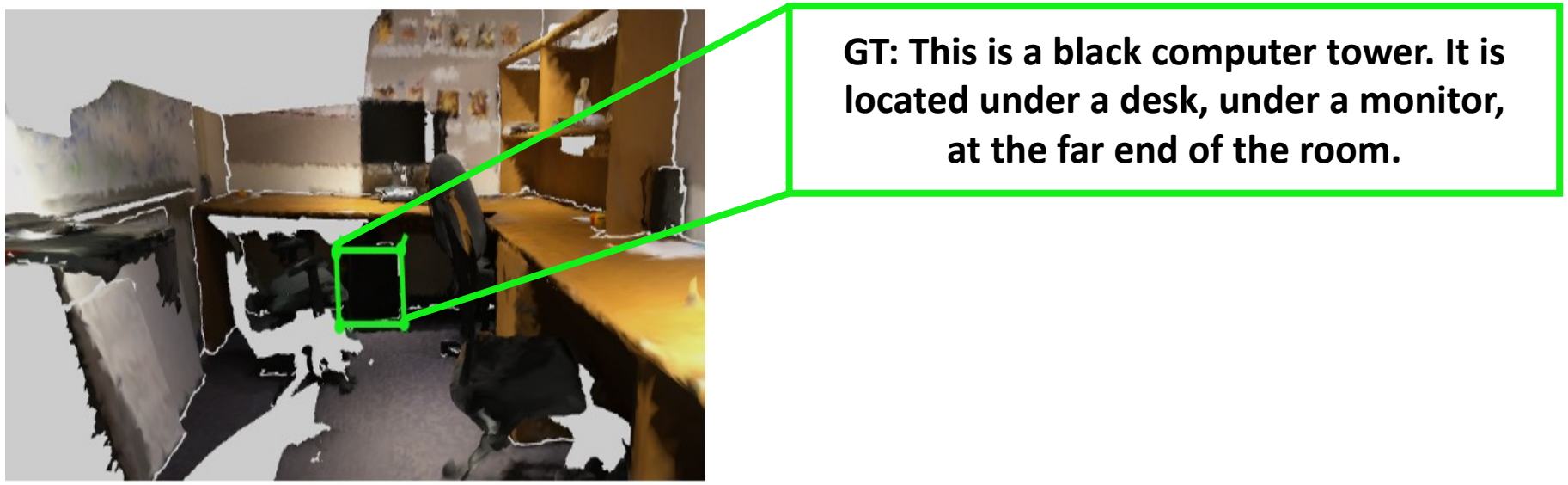
Minjung Kim, Hyung Suk Lim, Soonyoung Lee, Bumsoo Kim†, Gunhee Kim†

SEOUL NATIONAL UNIV.
VISION & LEARNING
LG AI Research

ECCV
EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO
2024

3D Dense Captioning

- 3D dense captioning is an advanced vision-language task that aims to generate multiple detailed and accurate descriptions for 3D scenes.
- This process typically involves identifying and localizing objects within a 3D point cloud and describing their attributes and spatial relationships.



Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- **Where should we look to correctly generate this caption?**

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- However, current Transformer Encoder-Decoder architectures **allocate a single query to a single object that needs to generate all the captions**.

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Motivation

- However, current Transformer Encoder-Decoder architectures allocate a single query to a single object that needs to generate all the captions.

This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Focusing the attention on local regions

➔ Better localization, individual object attribute detection

➔ Less contextual information, suboptimal relationship classification

Motivation

- However, current Transformer Encoder-Decoder architectures allocate a single query to a single object that needs to generate all the captions.

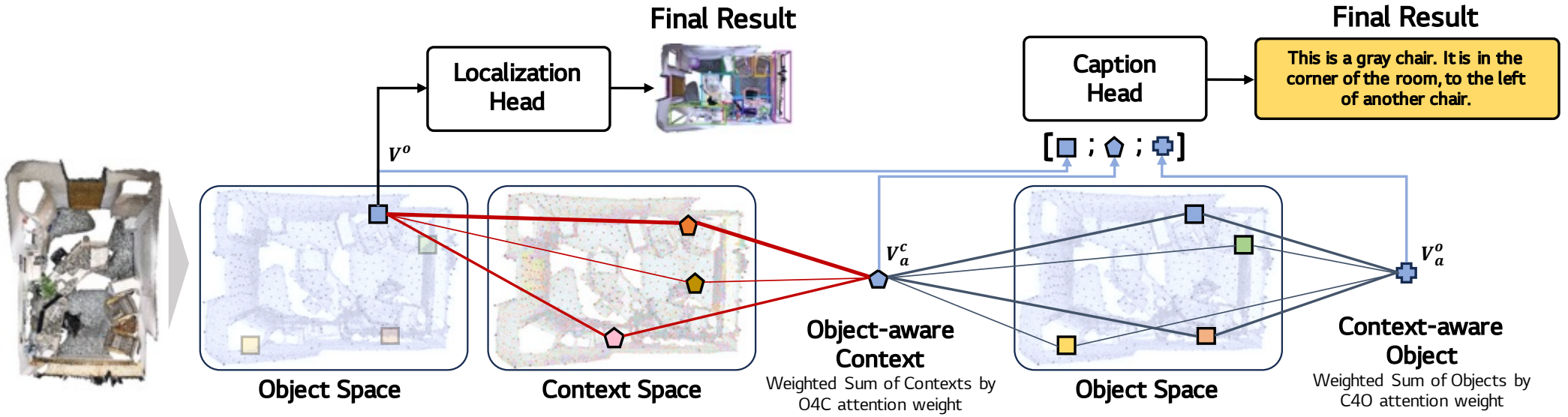
This is a wooden chair placed under the desk.
It is located right next to the window and
furthest from the door next to the bookshelf.

Contextualizing attention

➔ Better context recognition, better relationship caption generation

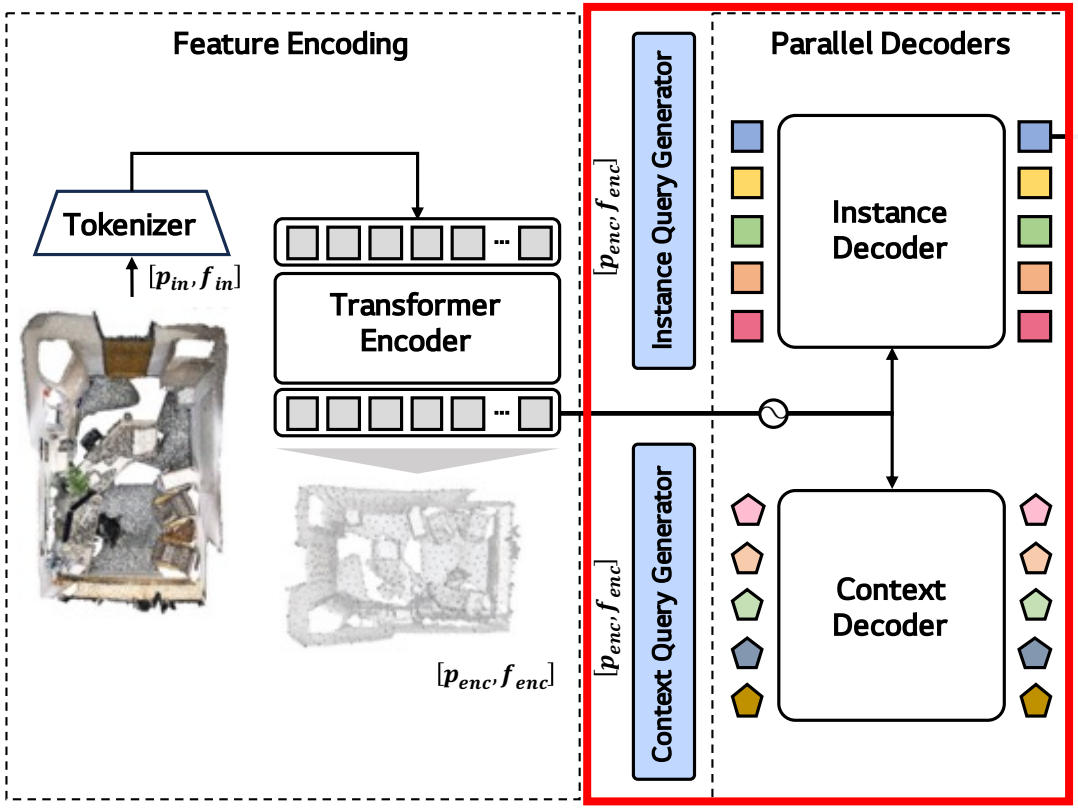
➔ Lower localization performance, sparse attention for individuals

Overview



Bi-directional **C**ontextual **A**ttention ➔ **BiCA**

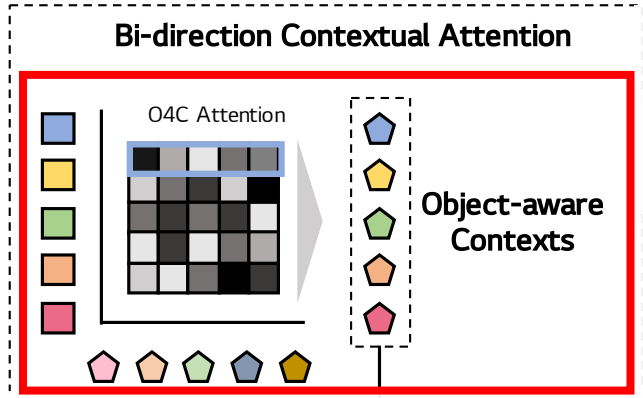
BiCA



We independently generate context and instance queries separately.

The instance decoder decodes a set of objects (■■■■■) while the context decoder decodes a set of context queries that can combine and contextualize the objects (⬟⬟⬟⬟⬟).

BiCA



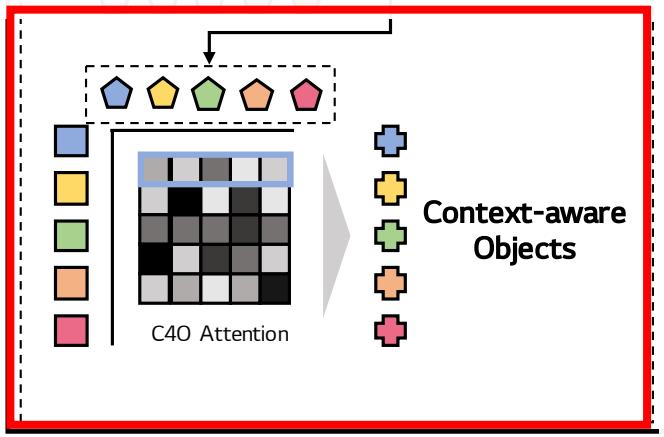
For each detected object, we then get the attention score with the ”**contexts**” in the scene.

This shows an **abstract contextualization** for which relation context position the current object of interest is participating.

BiCA

Finally, for each abstract contextualization for each object, we calculate the attention for individual instances.

As context queries represent the positional context feature for the relations involving a single object (■), the **contextualized object feature** aggregates the feature of all the participating objects according to their attention weight.



Experiments

- **ScanRefer dataset:**
  - The descriptions in the ScanRefer include depictions of the target object's attributes and information about the spatial relationships between this target object and other surrounding objects within the same space.

Model	Training	w/o additional 2D data								w/ additional 2D data							
		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	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
D3Net	MLE	-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
SpaCap3d	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
MORE	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
3DJCG	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
Contextual	MLE	-	-	-	-	42.77	23.60	22.05	45.13	-	-	-	-	46.11	25.47	22.64	45.96
REMAN	MLE	-	-	-	-	-	-	-	-	62.01	36.37	27.76	56.25	45.00	26.31	22.67	46.96
3D-VLP	MLE	64.09	39.84	27.65	58.78	50.02	31.87	24.53	51.17	70.73	41.03	28.14	59.72	54.94	32.31	24.83	51.51
Vote2Cap-DETR	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	52.38
Unit3D	MLE	-	-	-	-	-	-	-	-	-	-	-	-	46.69	27.22	21.91	45.98
Vote2Cap-DETR++	MLE	76.36	41.37	28.70	60.00	67.58	37.05	26.89	55.64	77.03	40.99	28.53	59.59	64.32	34.73	26.04	53.67
BiCA (Ours)	MLE	**78.42**	**41.46**	**28.82**	**60.02**	**68.46**	**38.23**	**27.56**	**58.56**	**78.35**	**41.20**	**28.82**	**59.80**	**66.47**	**36.13**	**26.71**	**54.54**
Scan2Cap	SCST	-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net	SCST	-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	25.72	53.90
χ-Tran2Cap	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
Contextual	SCST	-	-	-	-	50.29	25.64	22.57	44.71	-	-	-	-	54.30	27.24	23.30	45.81
Vote2Cap-DETR	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
Vote2Cap-DETR++	SCST	88.28	44.07	28.75	59.89	78.16	39.72	26.94	55.52	88.56	43.30	28.64	59.19	74.44	37.18	26.20	53.30
BiCA (Ours)	SCST	**89.72**	**44.97**	**28.96**	**60.69**	**80.14**	**40.16**	**27.76**	**56.10**	**89.34**	**44.56**	**28.74**	**59.33**	**76.34**	**37.34**	**26.60**	**54.00**

C : CiDER
B-4: BLEU-4
M: METEOR
R : ROUGH-L

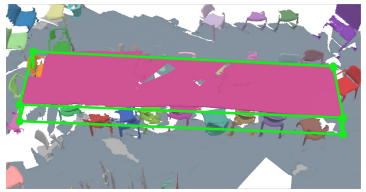
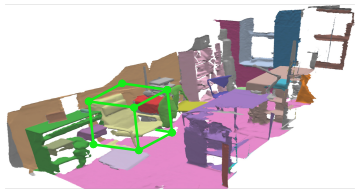
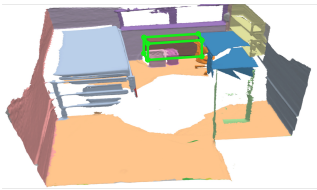
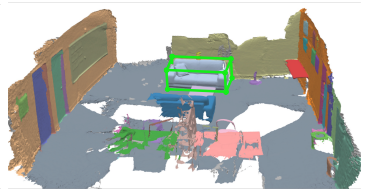
Experiments

- **Nr3D dataset:**
  - The Nr3D dataset is designed to evaluate the model's performance in interpreting free-form natural language descriptions of objects as spoken by humans.

Model	Training	C@0.5 ↑	B-4@0.5 ↑	M@0.5 ↑	R@0.5 ↑
Scan2Cap	MLE	27.47	17.24	21.80	49.06
D3Net		33.85	20.70	23.13	53.38
SpaCap3d		33.71	19.92	22.61	50.50
3DJCG		38.06	22.82	23.77	52.99
Contextual		35.26	20.42	22.77	50.78
REMAN		34.81	20.37	23.01	50.99
Vote2Cap-DETR		43.84	26.68	25.41	54.43
Vote2Cap-DETR++		47.08	27.70	25.44	55.22
BiCA (Ours)		**48.77**	**28.35**	**25.60**	**55.81**
D3Net	SCST	38.42	22.22	24.74	54.37
χ-Tran2Cap		33.62	19.29	22.27	50.00
Contextual		37.37	20.96	22.89	51.11
Vote2Cap-DETR		45.53	26.88	25.43	54.76
Vote2Cap-DETR++		47.62	28.41	25.63	54.77
BiCA (Ours)		**49.81**	**28.83**	**25.85**	**56.46**

C : CiDER
B-4: BLEU-4
M: METEOR
R : ROUGH-L

Qualitative Results

Scene0249_00|24|table	*Scene0535_00|15|chair*	*Scene0277_00|12|radiator*	*Scene0329_00|1|couch*
			
SpaCap3D : FTG.	**SpaCap3D :** This is a brown chair. It is to the right of a white chair.	**SpaCap3D :** The radiator is white and short. The radiator is on the wall between two windows.	**SpaCap3D :** This is a couch. Its brown and blue and white in color and is next to a table with a lamp on top of it. Its located.
3DJCG : FTG.	**3DJCG :** This is a black chair. It is to the left of the white table.	**3DJCG :** FTG.	**3DJCG :** The couch is the left of the couch. It is to the right of the couch.
Vote2Cap-DETR : The table is in the center of the room. It is to the left of the table.	**Vote2Cap-DETR :** This is a blue chair. It is to the right of the room.	**Vote2Cap-DETR :** This is a white radiator. It is to the right of the room.	**Vote2Cap-DETR :** The couch is on the right. It is to the right of the room.
Vote2Cap-DETR++ : FTG.	**Vote2Cap-DETR++ :** The chair is in the corner of the room. It is to the right of the table.	**Vote2Cap-DETR++ :** This is a white radiator. It is to the right of the desk.	**Vote2Cap-DETR++ :** The couch is on the right. It is to the right of the table.
Ours : This is a large rectangular table. It is surrounded by chairs.	**Ours :** The chair is the northwest – most one in the room. The chair is black and has four legs.	**Ours :** This is white radiator. It is under the window.	**Ours :** The couch is the rightmost one in the room. The couch is a dark brown rectangle.
GT : The table is in the center of the room. The table is a long rectangle.	**GT :** The chair is against the wall in the back of the room. Its grey and black and has wooden arms and sits next to a desk.	**GT :** It is radiator under a window. The radiator is to the right of the bed.	**GT :** There is a brown leather couch. Placed on the side of the wall.

: Object itself

: The spatial position of the object in the 3D scene

: Relationships between objects.

FTG. : Failures in caption generation due to low IoU

Thank you!

LinkedIn

Project page