Technische
Universität
München

SoftCap: Dense Captioning for 3D Scenes with SparseConv

Yunxiang Lu, Jiachen Lu

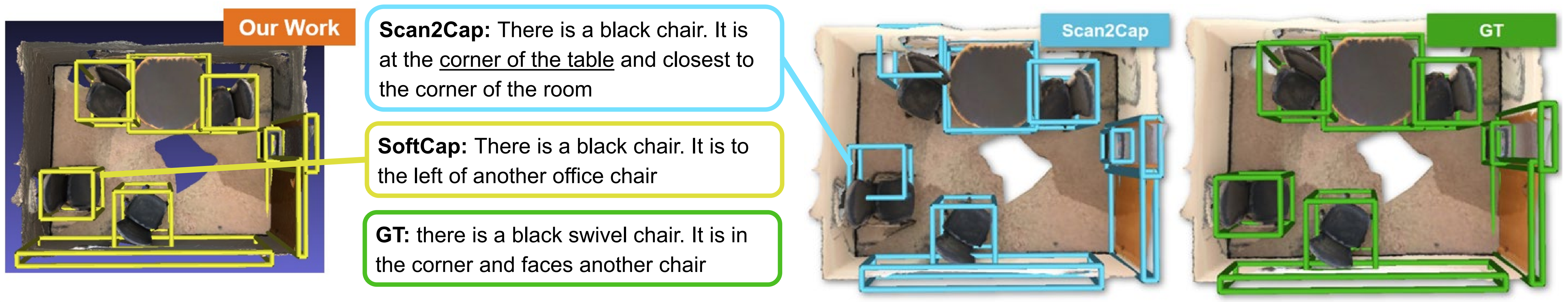
Introduction

Recent works on 3D dense captioning have achieved impressive results. However, most existing methods such as Scan2Cap, X-TransCap and MORE all use VoteNet as the detection backbone. The limited performance of object detection constrains the quality of generated captions.

To address this issue, we propose a model using **SoftGroup based detection backbone**. With sparse convolution and soft grouping mechanism, better detection performance and denser object features can be achieved, which enables the later language model to generate more reliable captions. A relational graph module and a Context-aware Attention Captioning module are used to aggregate object features with relational information. Our method can effectively localize and describe objects in 3D scenes and outperforms the existing baseline method.

Our main works can be summarized as following:

- We propose an End-to-End 3D dense captioning model SoftCap with a more powerful SoftGroup based detection backbone compared to existing methods, showing that denser object features contributes to more reliable descriptions.
- We adapt the self-critical sequence training mechanism in 3D dense captioning task, showing that “REINFORCE with baseline” algorithm further improves the captioning performance.
- We study how the different components in the model impact the final captioning performance based on our detection backbone.



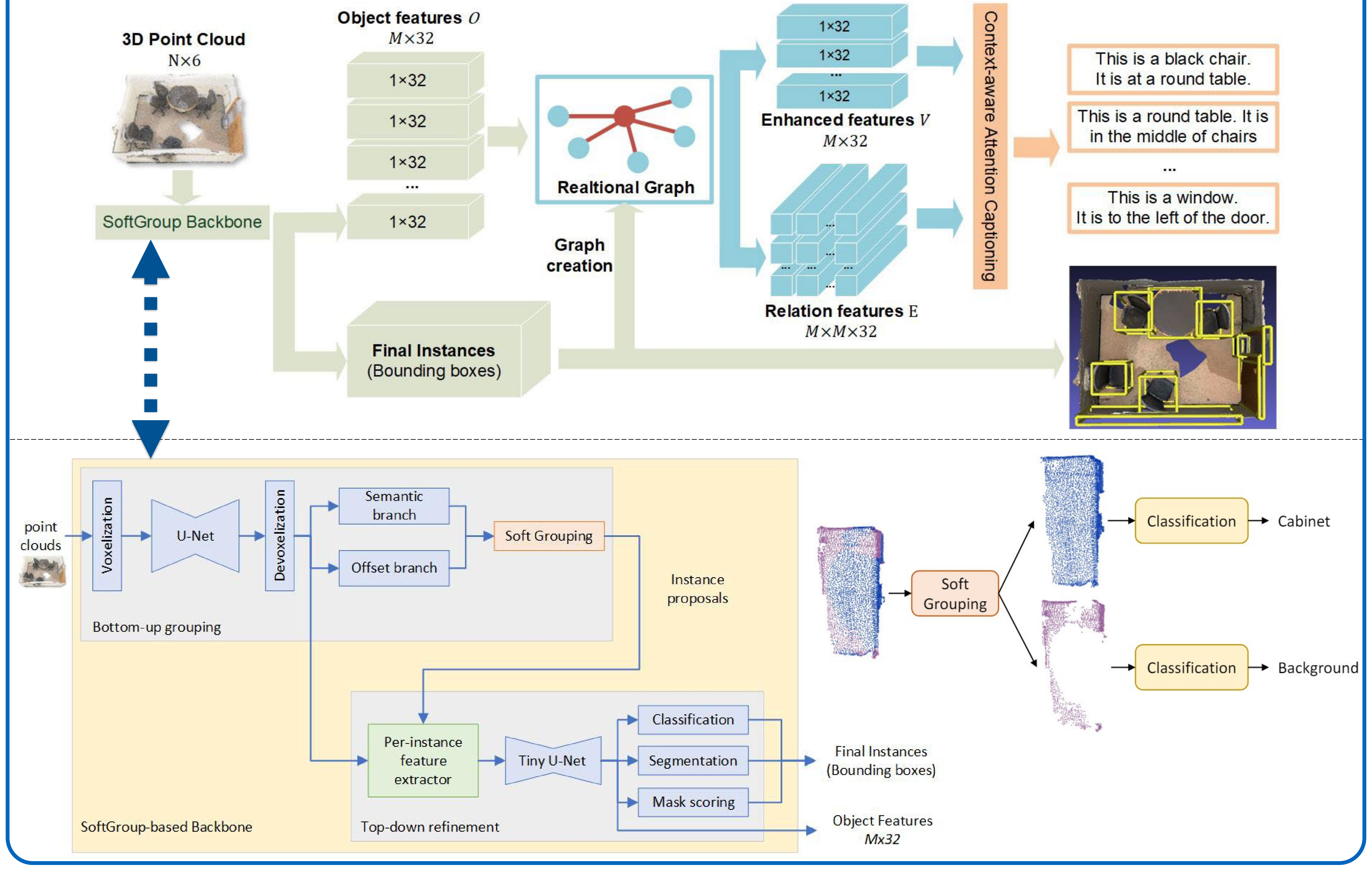
Scan2Cap: There is a black chair. It is at the corner of the table and closest to the corner of the room

SoftCap: There is a black chair. It is to the left of another office chair

GT: there is a black swivel chair. It is in the corner and faces another chair

End-to-End SoftCap Network Architecture

- SoftGroup-based Detection Backbone → Detect potential objects in 3D scene
- Relational Graph → Enhance object features and extract relation features
- Context-aware Attention Captioning → Aggregate features and generate descriptions



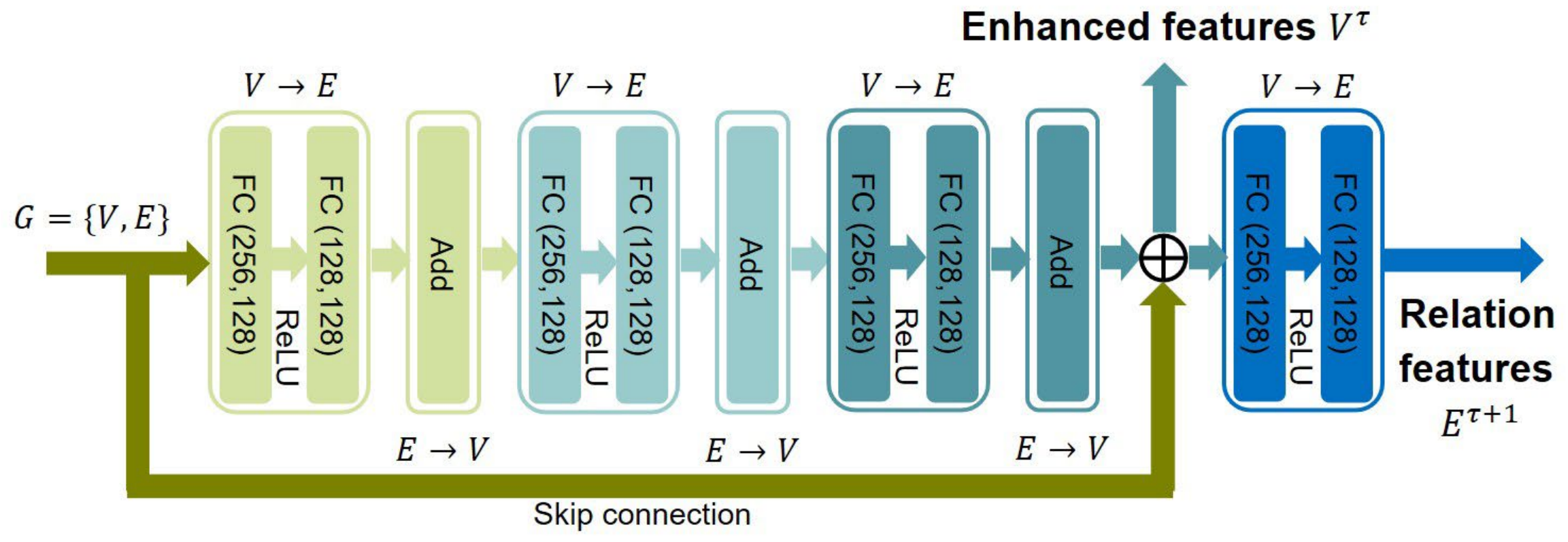
Method

SoftGroup-based Detection Backbone:

The detection backbone consists of bottom-up grouping and top-down refinement stages. **In the grouping stage**, the grouping process is performed based on the predicted soft semantic scores and offset vectors. **In the refinement stage**, the corresponding proposal features are extracted and used to predict classes, instance masks, and mask scores for getting final instances.

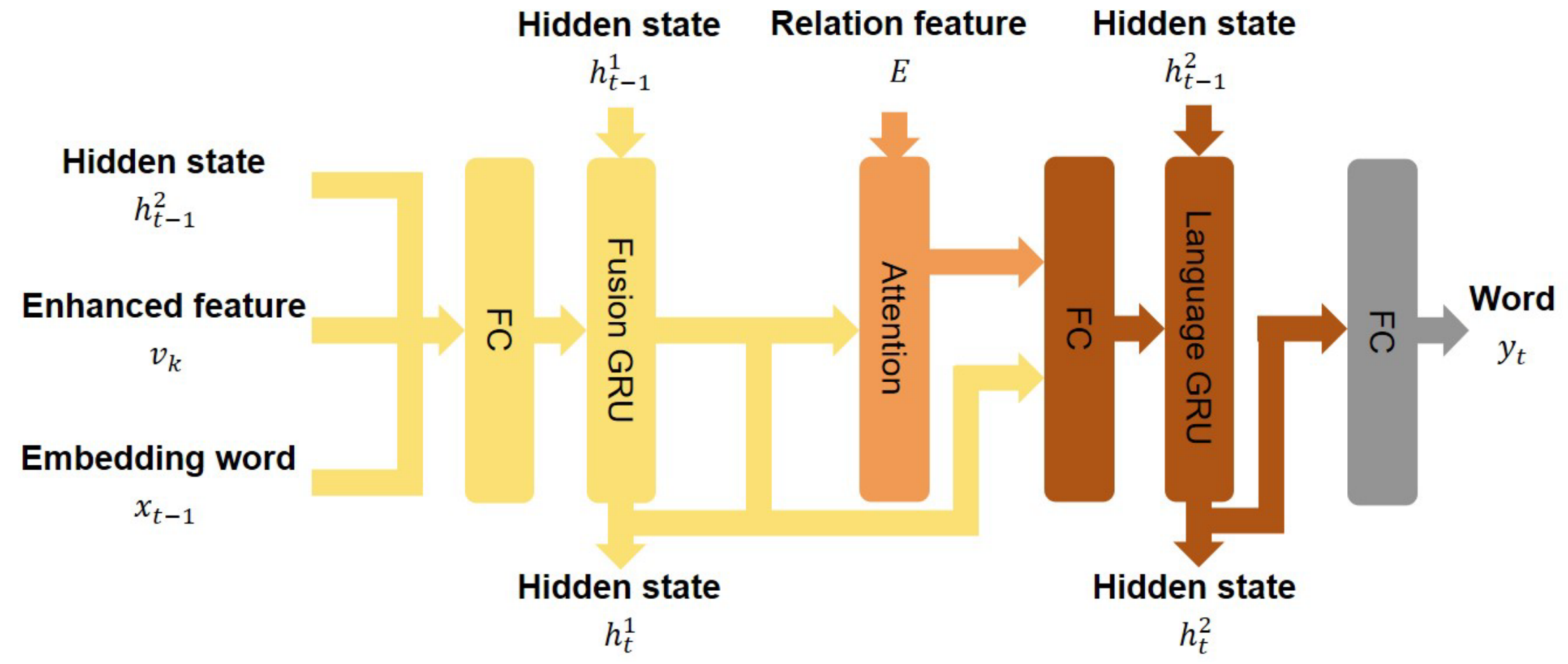
Relational Graph (RG):

We use predicted instance bounding boxes to create a relational graph module $G = (V, E)$ equipped with a message passing network to enhance the object features and extract the object relation features. Each node representing an object is connected to its K nearest neighbors in the graph.



Context-aware Attention Captioning (CAC):

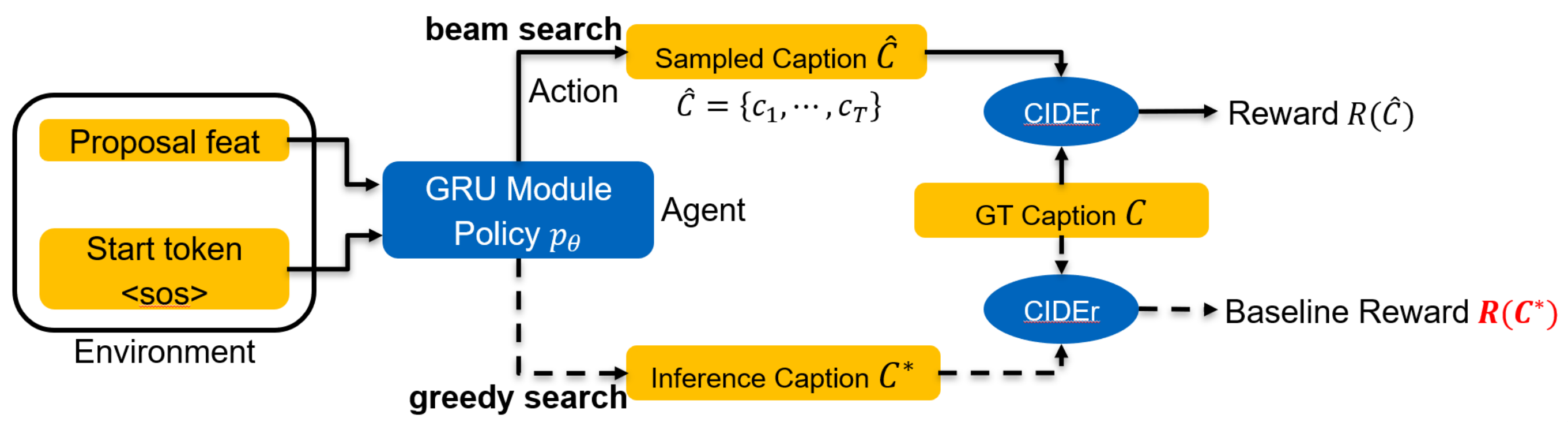
The CAC language module takes both the enhanced object features V and object relation features E and generate the caption one token at a time. It uses a Fusion GRU layer and an Attention Layer to aggregate the final context features representing the attended objects and inter-object relations.



REINFORCE with baseline

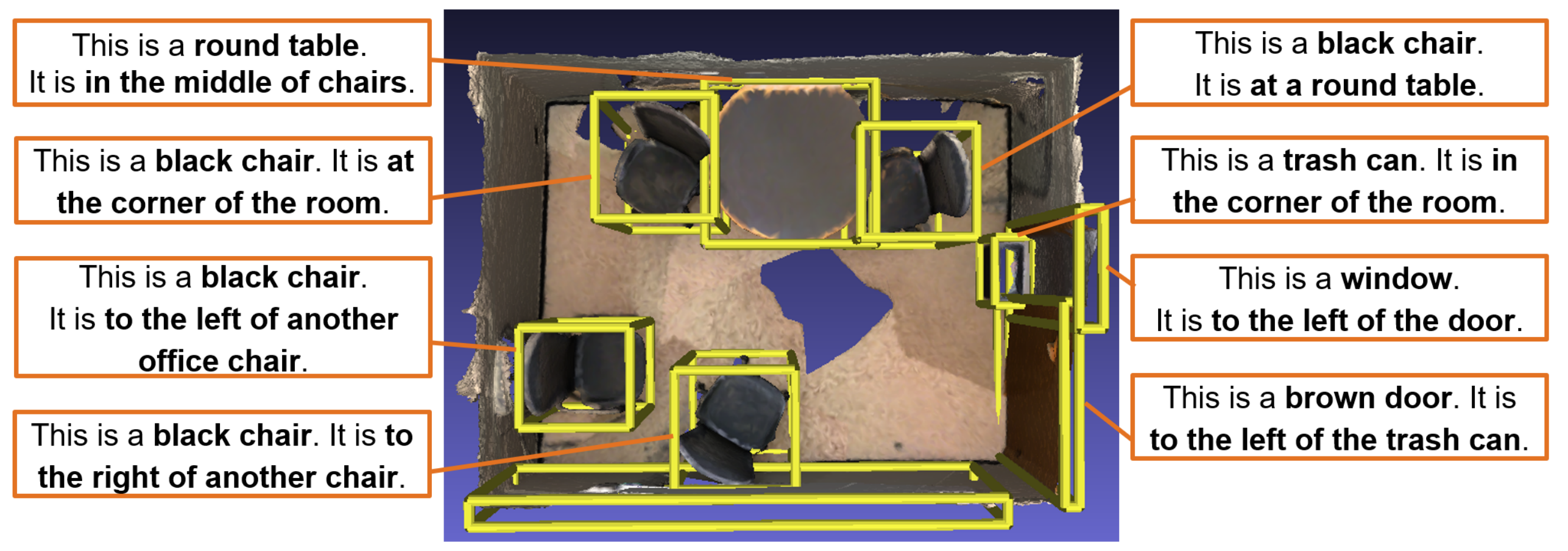
We apply “REINFORCE with baseline” algorithm to further improve the language module. The score of inference caption C^* defines the baseline reward and we use policy gradient to construct final loss function:

$$L_{cap}(\theta) \approx -\left[R(\hat{C}) - \boldsymbol{R(C^*)}\right] \sum_{t=1}^{T} \log p(\hat{c}_t|\theta)$$



Training and Inference

Pre-train SoftGroup → Train SoftCap with CE → Train SoftCap with REINFORCE



This is a **round table**. It is **in the middle of chairs**.

This is a **black chair**. It is **at the corner of the room**.

This is a **black chair**. It is **to the left of another office chair**.

This is a **black chair**. It is **to the right of another chair**.

This is a **black chair**. It is **at a round table**.

This is a **trash can**. It is **in the corner of the room**.

This is a **window**. It is **to the left of the door**.

This is a **brown door**. It is **to the left of the trash can**.

Inference result on ScanRefer Val after training with CE

Experimental Results

- Quantitative Analysis

		Captioning F1-Score @0.5IoU				Detection @0.5IoU
Method	**Detection**	**C**	**B-4**	**M**	**R**	**mAP**
Scan2Cap	VoteNet	15.71	9.01	7.18	14.92	32.09
X-Trans2Cap	VoteNet	17.64	9.68	7.21	15.25	35.31
MORE	VoteNet	16.46	8.86	7.12	14.71	31.93
Ours (CE Loss)	SoftGroup	30.76	16.30	**13.83**	28.41	57.22
Ours (CIDEr Loss)	SoftGroup	**36.27**	**18.66**	13.82	**29.13**	**57.38**

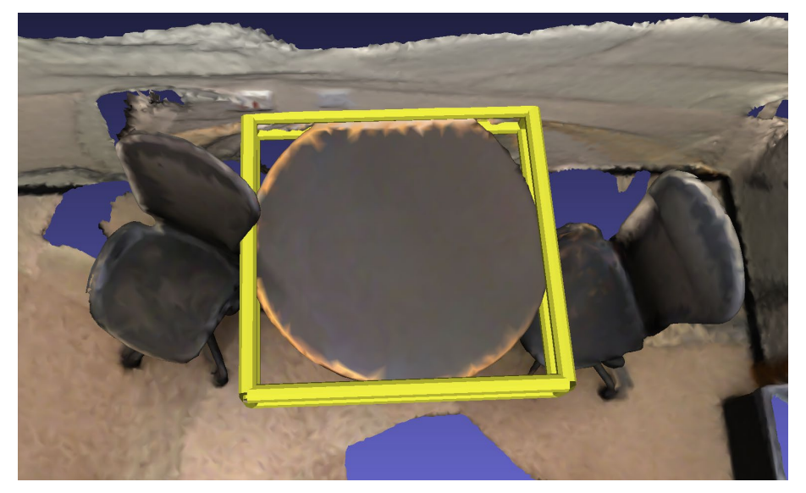
- Ablation Study 1 (CE Loss)

	Captioning F1-Score @0.5IoU				Detection @0.5IoU
Network Architecture	**C**	**B-4**	**M**	**R**	**mAP**
SoftGroup + GRU	25.69	13.74	12.96	26.88	55.64
SoftGroup + RG + GRU	26.12	14.09	12.74	26.98	55.13
SoftGroup + RG + Att2GRU	26.77	14.81	13.13	27.48	56.48
SoftGroup + RG + CAC	**30.76**	**16.30**	**13.83**	**28.41**	**57.22**

- Ablation Study 2 (CIDEr Loss)

	Captioning F1-Score @0.5IoU				Detection @0.5IoU
Network Architecture	**C**	**B-4**	**M**	**R**	**mAP**
SoftGroup + GRU	33.24	17.45	13.57	28.36	55.80
SoftGroup + RG + GRU	34.12	17.38	13.62	28.61	55.29
SoftGroup + RG + Att2GRU	34.78	17.62	13.46	28.23	56.64
SoftGroup + RG + CAC	**36.27**	**18.66**	**13.82**	**29.13**	**57.38**

- Qualitative Analysis

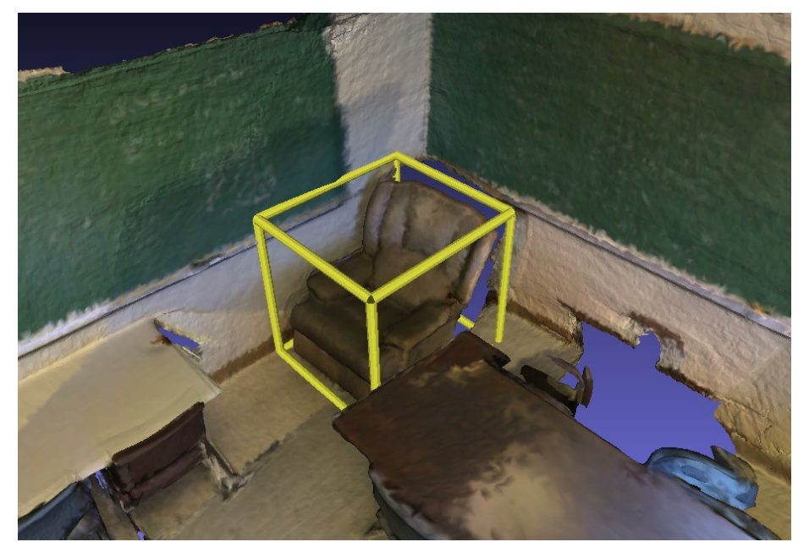


GT: The **round table** with the **black top** is **in the corner** of the room. **Two black office chairs around** the table.

SoftGroup + GRU: This is a **brown table**. It is **in the center of the room.**

SoftGroup + RG + CAC with CE: This is a **round table**. It is **in the middle of chairs.**

SoftGroup + RG + CAC with CIDEr: The table is a **round table.** It is **in the right of the room.**



GT: A **cushion sofa** in **brown color**. it is **near the wall**.

SoftGroup + GRU: This is a **brown chair**. It is **at the corner of the room**.

SoftGroup + RG + CAC with CE: The **arm-chair** is the **one closest to the window**. the armchair has **a curved backside** and **four legs.**

SoftGroup + RG + CAC with CIDEr: The chair is a **black chair**. it is **to the left of the table**.

Conclusions

In this work, we propose an End-to-End SoftGroup-based network SoftCap, which addresses the issue of unideal detection backbone in previous 3D dense captioning methods. Also, we adapt relational graph, CAC language module and “REINFORCE with baseline” algorithm to generate object descriptions. Our network outperforms the previous works in both detection and caption with significant improvement.

References

1. Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in RGB-D scans. CoRR, abs/2012.02206, 2020
2. Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds, 2022
3. Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. CoRR, abs/1612.00563, 2016.