Dense Captioning for 3D scenes with SparseConv

Alex Khakhlyuk and Anurag Singh
Supervisor: Dave Zhenyu Chen

Problem Definition:

- The task is dense captioning in 3D scans from commodity RGB-D sensors. As input, we assume a point cloud of a 3D scene; the expected output is the bounding boxes along with the descriptions for the underlying objects.

Our Contributions:

- We proposed to change the feature extraction backbone in Scan2Cap from PointNet++ to a SparseConv Unet.

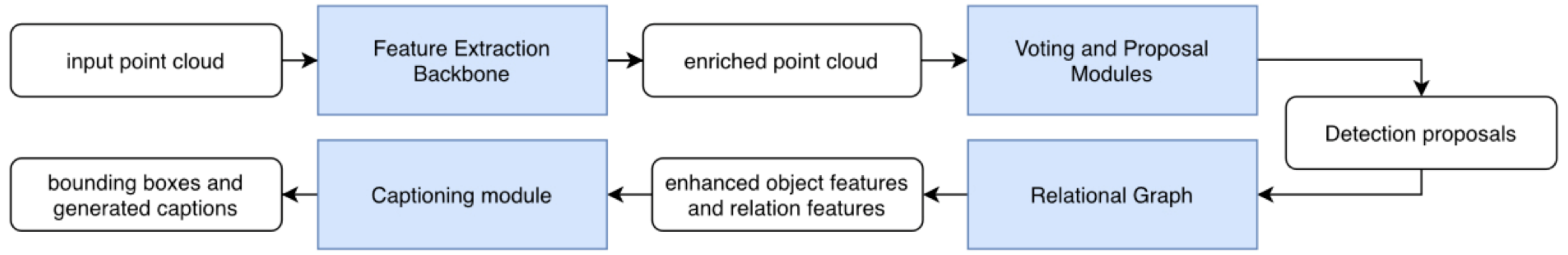

input point cloud
Feature Extraction Backbone
enriched point cloud
Voting and Proposal Modules
Detection proposals
Relational Graph
enhanced object features and relation features
Captioning module
bounding boxes and generated captions

Fig 1: Overview of Scan2Cap architecture (Chen et al CVPR 2021)

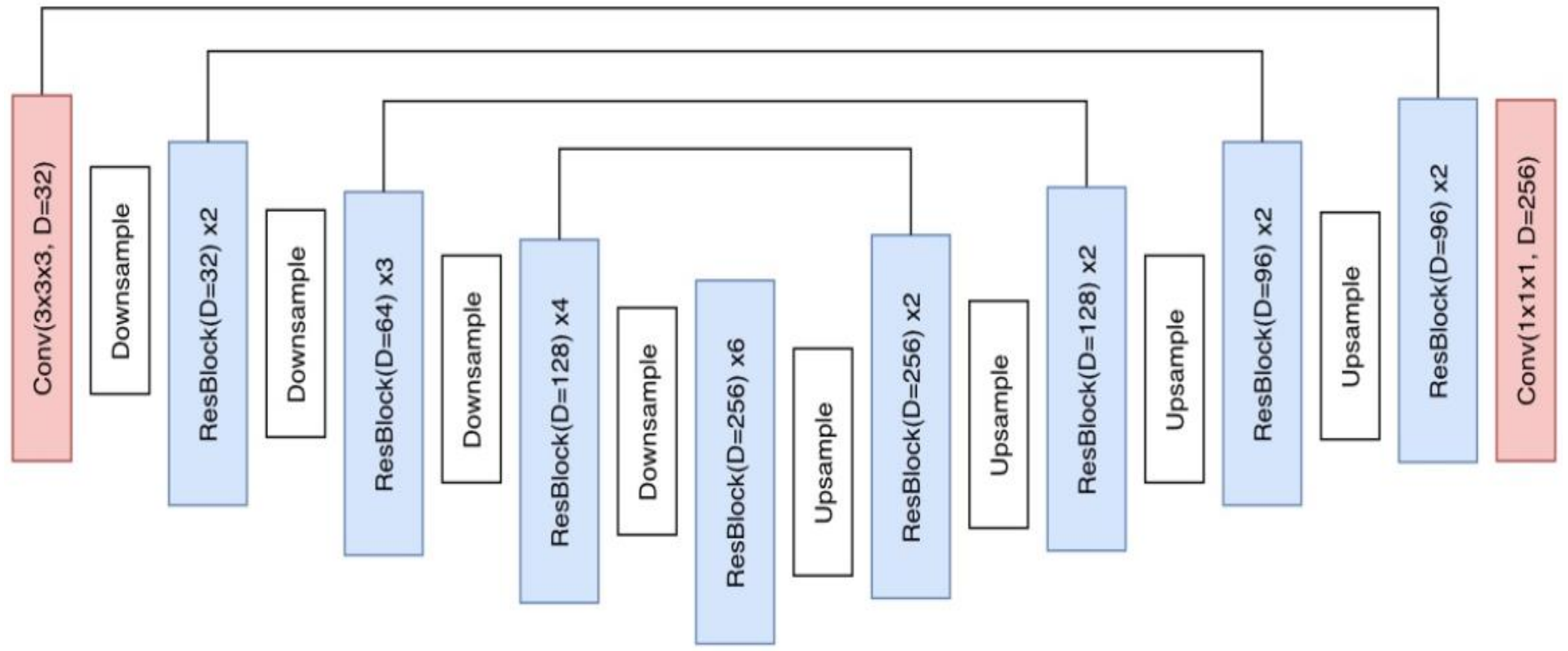


Fig 2: SparseUnet architecture for detection backbone in Scan2Cap

PointNet++

- Fast training and inference
- Lower accuracy

Sparseconv

- Slower training and inference
- High accuracy

Experiments:

- We evaluate detection and captioning performance of Scan2Cap with different backbones: PointNet++ vs SparseUnet.

backbone	mAP@0.25IoU	mAP@0.5IoU
PointNet++	51.64	28.80
SparseUnet	**52.05**	**33.59**

Table 1: Object detection.

backbone	B-4	C	M	R
PointNet++	31.54	44.59	25.06	**53.67**
SparseUnet	**32.30**	**49.52**	**25.52**	53.53

Table 2: Dense captioning with IoU@0.25

backbone	B-4	C	M	R
PointNet++	21.67	31.58	21.10	43.87
SparseUnet	**23.37**	**35.59**	**21.66**	**44.34**

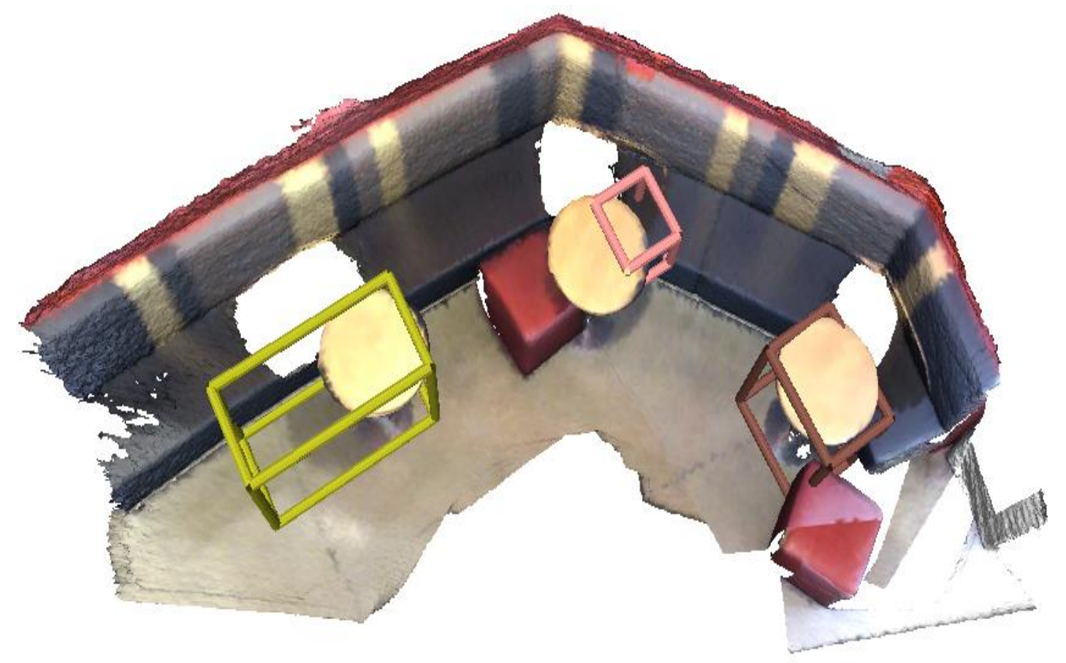
Table 3: Dense captioning with IoU@0.5

Task	Backbone	Memory	Forward	Forward+ Backward	Training time	Parameters
Detection	PointNet++	6.7GB	0.22s	0.9s	7h	1.0M
	SparseUnet	7.5 GB	0.82s	2.5s	23h	38M
Captioning	PointNet++	8.0 GB	0.82s	1.4s	39h	2.7M
	SparseUnet	8.8GB	1.15s	3s	70h	40M

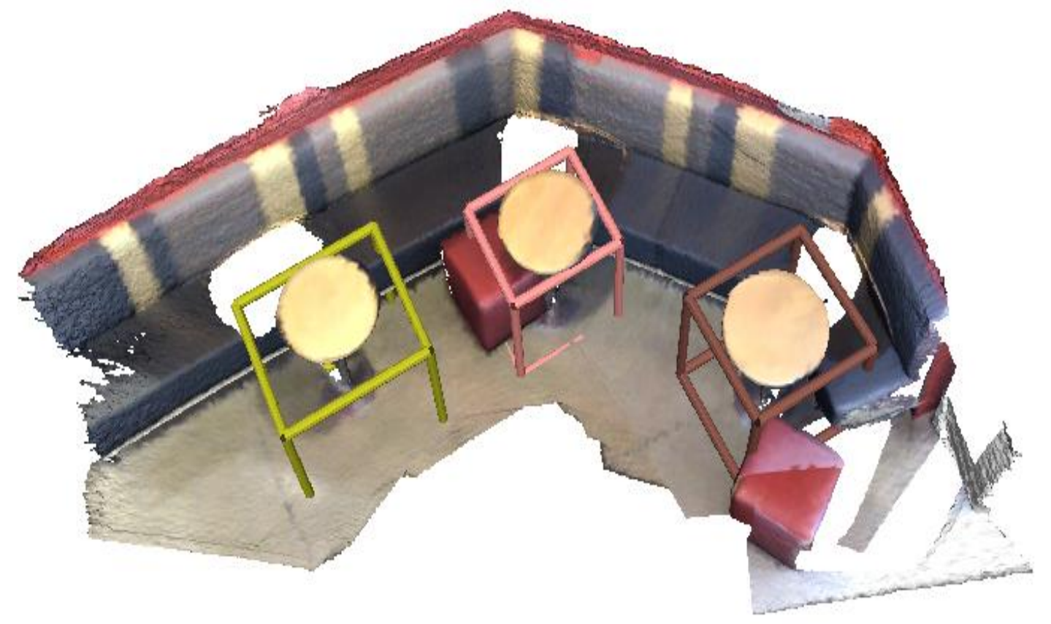
Table 4: Comparison of time and memory requirements for both tasks.

Qualitative Visualizations (Detection):

- SparseUnet produces more accurate bounding boxes.



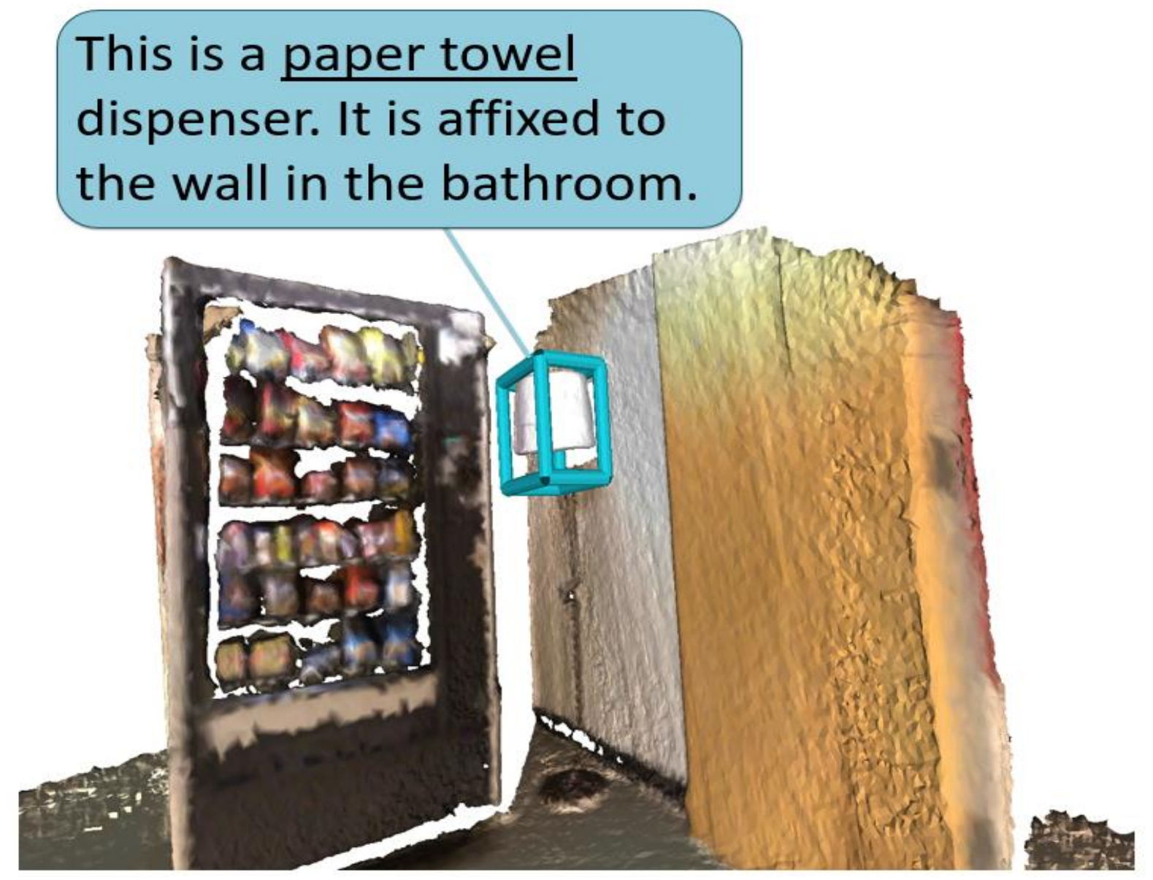
Pointnet++



SparseUnet

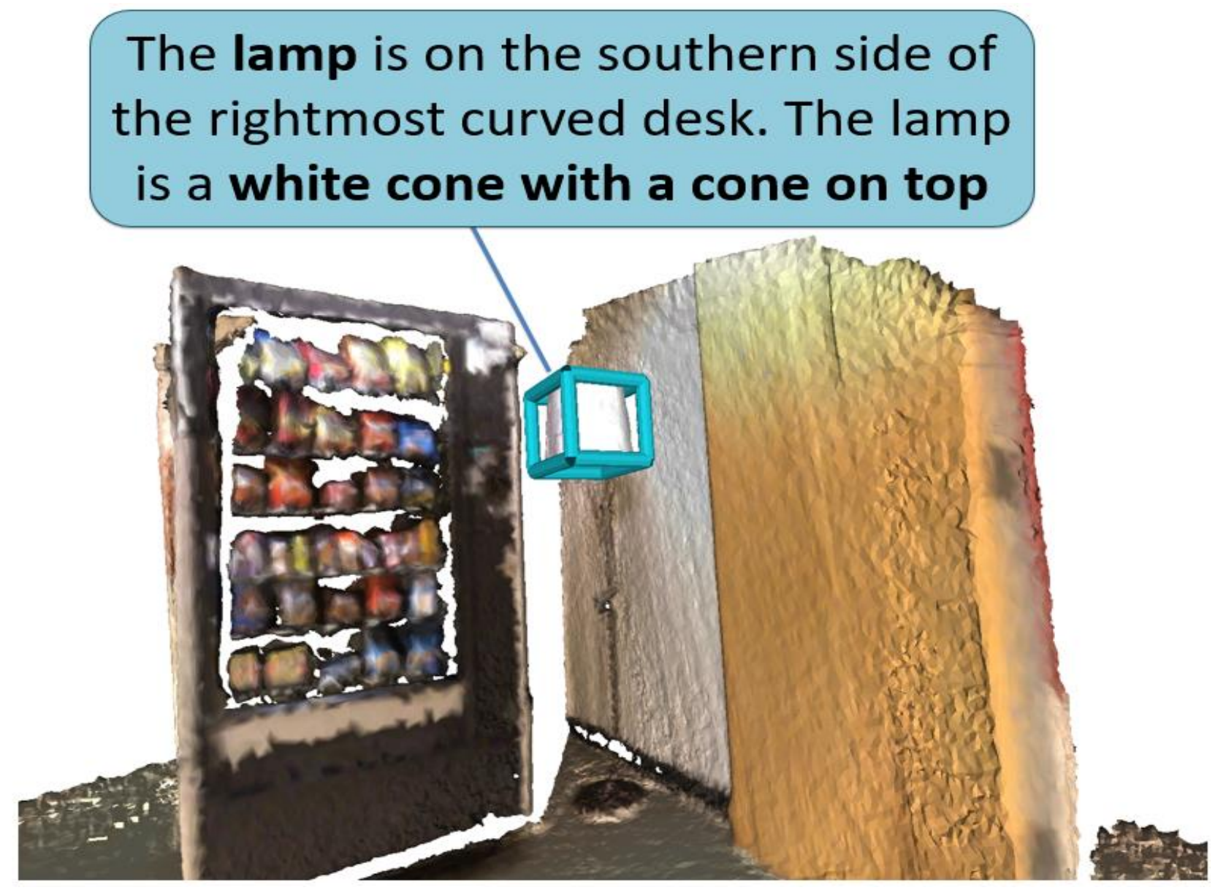
Qualitative Visualizations (Dense Captioning):

- Due to better features and detection proposals, SparseUnet generates accurate captions with correct classes.



This is a paper towel dispenser. It is affixed to the wall in the bathroom.

Pointnet++

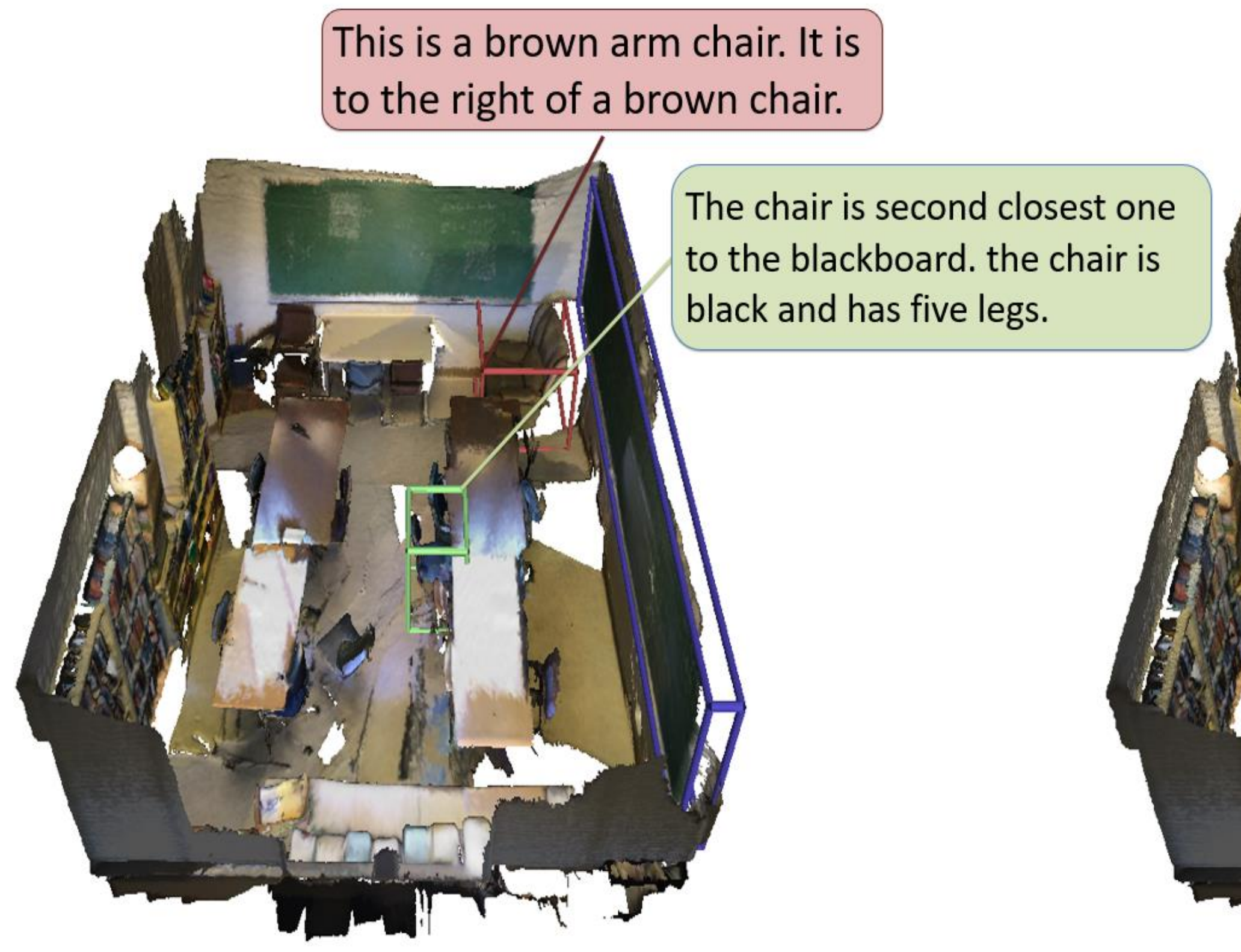


The **lamp** is on the southern side of the rightmost curved desk. The lamp is a **white cone with a cone on top**

SparseUnet

.

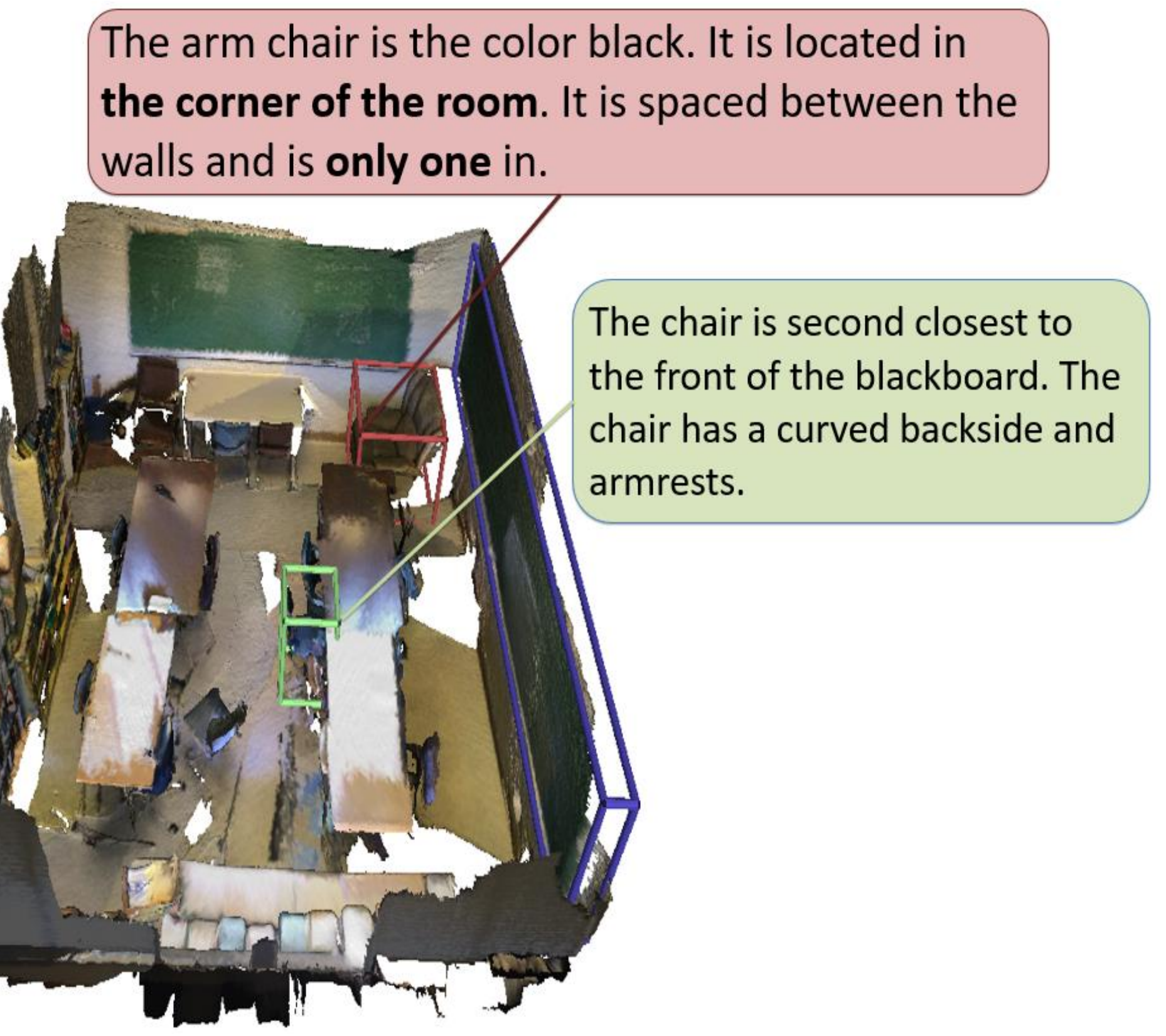
- We observe that SparseUnet is able to incorporate global semantics better into the captions.



This is a brown arm chair. It is to the right of a brown chair.

The chair is second closest one to the blackboard. the chair is black and has five legs.

Pointnet++



The arm chair is the color black. It is located in **the corner of the room**. It is spaced between the walls and is **only one** in.

The chair is second closest to the front of the blackboard. The chair has a curved backside and armrests.

SparseUnet

Summary/Conclusion

- SparseUnet results in improvement of performance both for object detection and dense captioning.
- SparseUnet is better at capturing global semantics and object location.
- The performance boost from SparseUnet comes at a cost of training and inference speed.