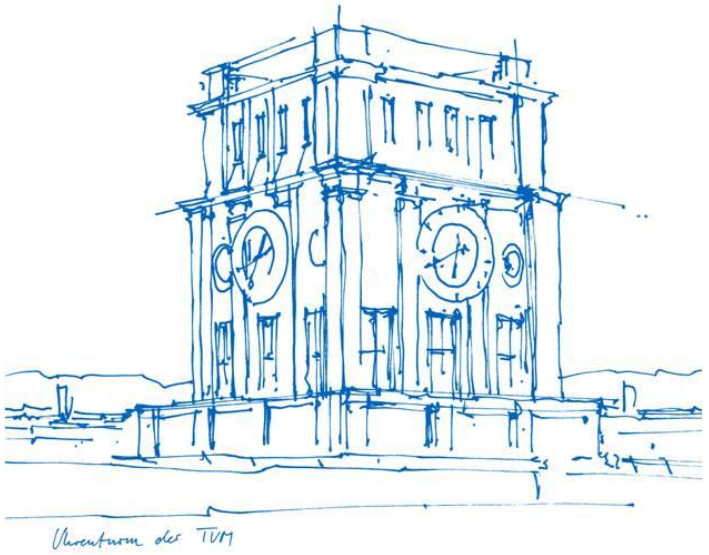
3D Visual Grounding with Transformers

2nd Presentation

Advanced Deep Learning for Computer Vision (IN2364)

Department of Computer Science

Technical University of Munich



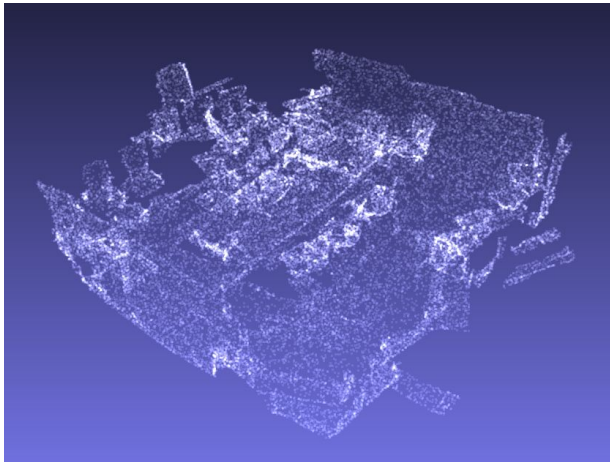
Bastian Wittmann, Philipp Foth

Munich, 28.06.21

Agenda

1. Recap
2. Experiments with different loss functions
3. Results
4. Visualizations
5. More input features
6. Open challenges and next steps

Recap - Visual Grounding



Input:

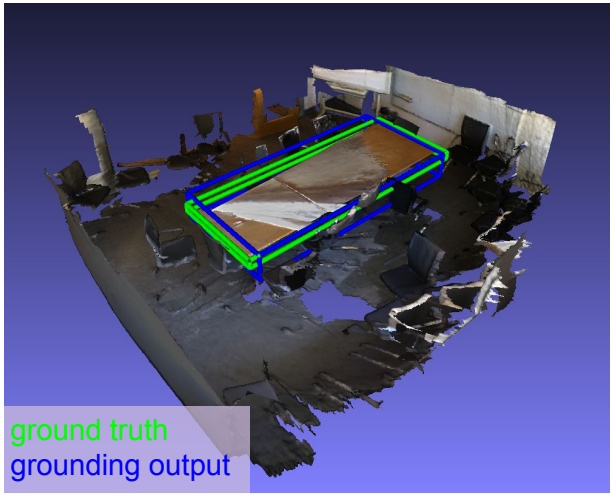
1) point cloud

2) natural language description

'this is a long bare table. it is in the middle of many chairs.'

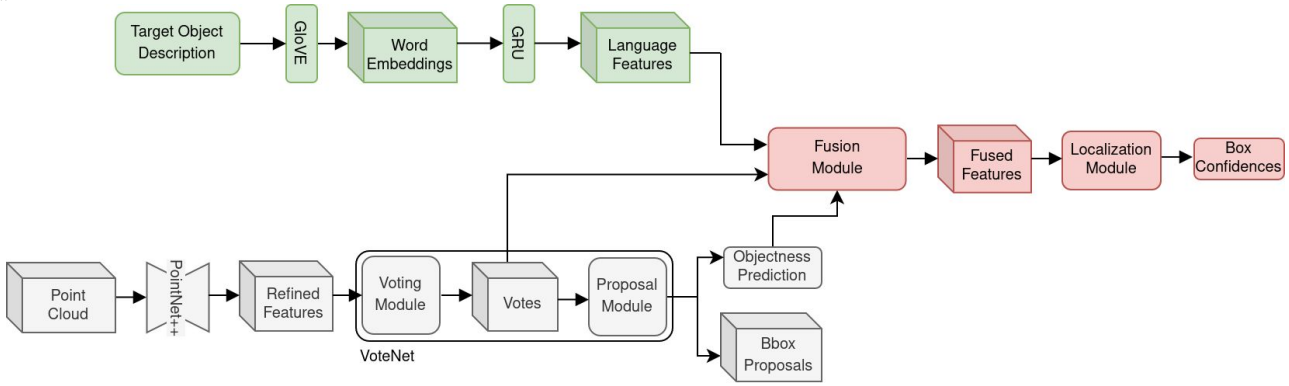
Output:

bounding box corresponding to the object described in the natural language description

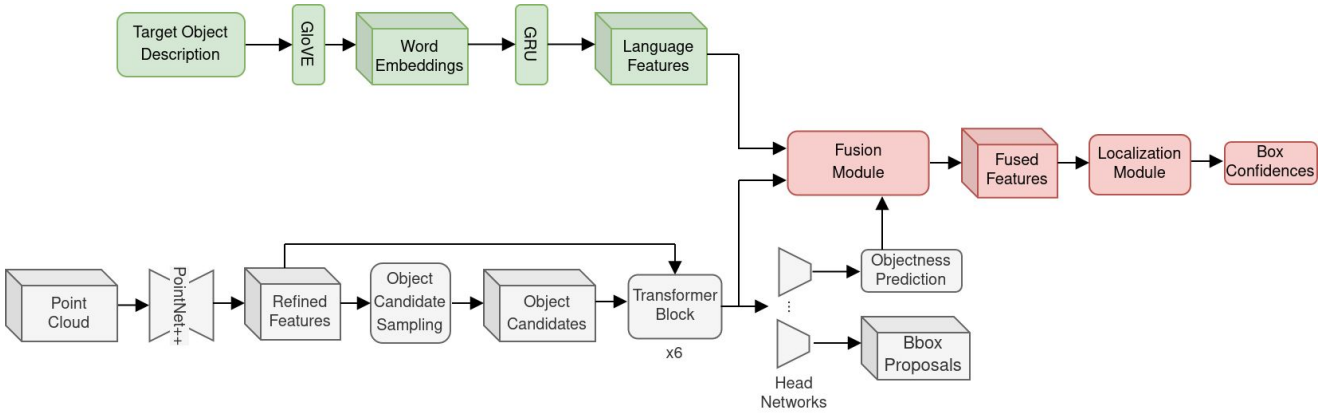


Recap - RefNet[1] vs. RefNetV2

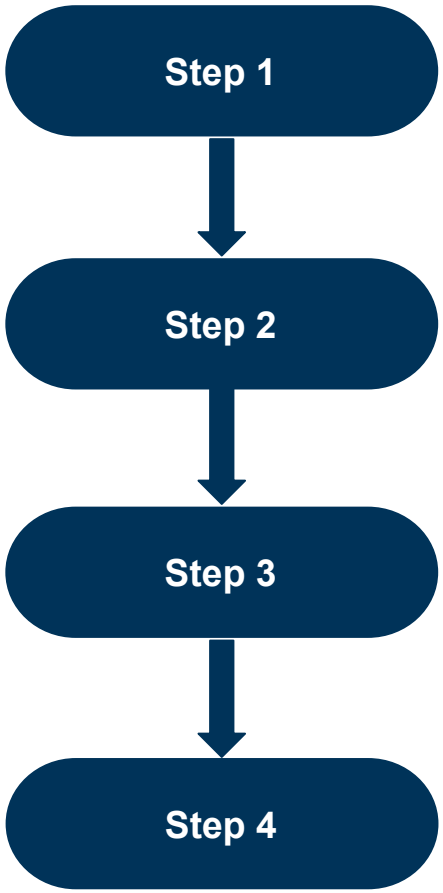
RefNet "Vanilla RefNet"



RefNetV2



Recap - Progress So Far



- Setup infrastructure
- Validate claims

- Replace VoteNet
- Overfit on a few scenes
- Train using frozen pre-trained detector
- Gradually unfreeze layers of the detector

- Experiments with different loss functions
- Improve performance via hyperparameter search
- Improve performance with more input features

- Submit to the ScanRefer Benchmark Challenge
- Experiments with contrastive loss

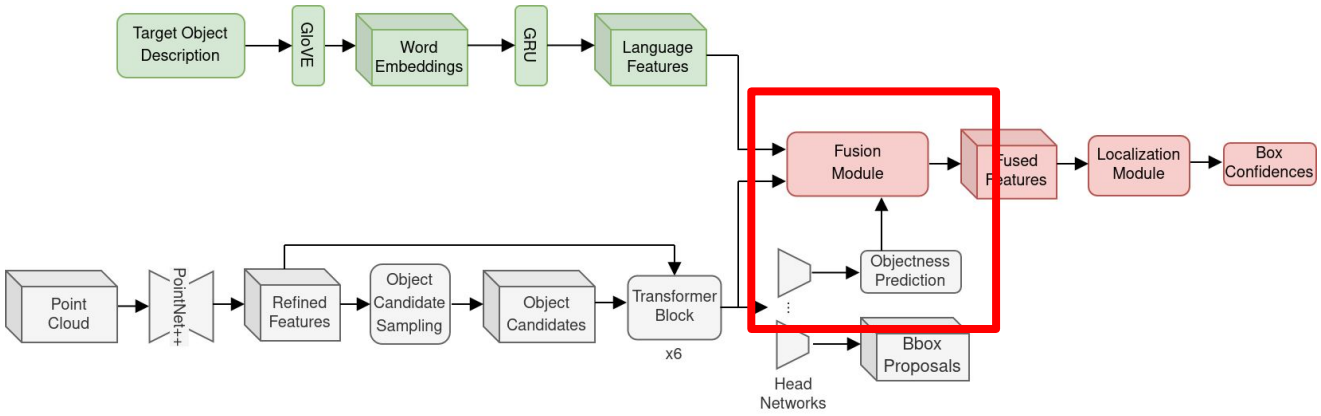
Concerns with New Detection Module

RefNetV2 during training has lower:

- Objectness accuracy
- Reference accuracy

Why?

Problematic?

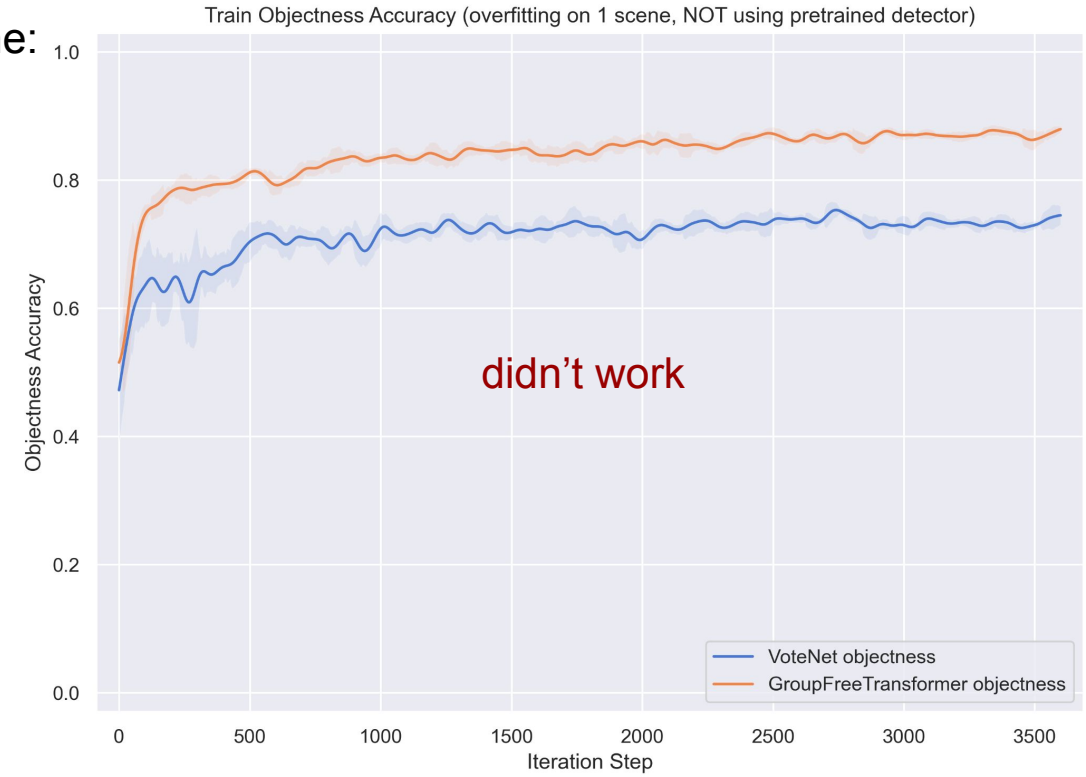


Objectness Accuracy

Different objectness losses

Adapt transformer-detector with VoteNet-detector objectness loss

Train on one scene:



Reference Ground Truth

Ground truth (GT) assigned dynamically:

predicted bounding boxes (bboxes):

$$bbox_{pred} = [bbox1, bbox2, bbox3, bbox4 \ldots, bboxN]$$

IoU (overlap) with reference GT bbox:

$$IoU_{pred\_bbox,\ ref\_gt\_bbox} = [0, 0, 0.8, 0.2, \ldots, 0]$$

max IoU predicted bbox gets assigned positive GT:

$$ref_{gt} = [0, 0, 1, 0, \ldots, 0]$$

Multi Reference Ground Truth

$$IoU_{pred\_bbox,\ ref\_gt\_bbox} = [0.81, 0, 0.8, 0.1, \ldots, 0.79]$$

Problem with “**single reference**”:

max IoU predicted bbox gets assigned positive GT:

$$ref_{gt} = [1, 0, 0, 0, \ldots, 0]$$

Propose “**multi reference**” ground truth:

IoU predicted bbox > THRESHOLD (0.3) get assigned positive GT:

$$ref_{gt} = [1, 0, 1, 0, \ldots, 1]$$

Reference Prediction

$IoU_{pred\_bbox,\ ref\_gt\_bbox}$ = [0.82, 0, 0.8, 0.2, …, 0]

Single reference:

ref_{gt} = [1, 0, 0, 0, …, 0]

Prediction:

object candidate features

↓

mask with objectness predictions

↓

logits

↓

softmax

ref_{pred} = [0.8, 0, 0.1, 0.1, …, 0]

Multi reference:

ref_{gt} = [1, 0, 1, 0, …, 0]

Prediction:

object candidate features

↓

logits

↓

sigmoid

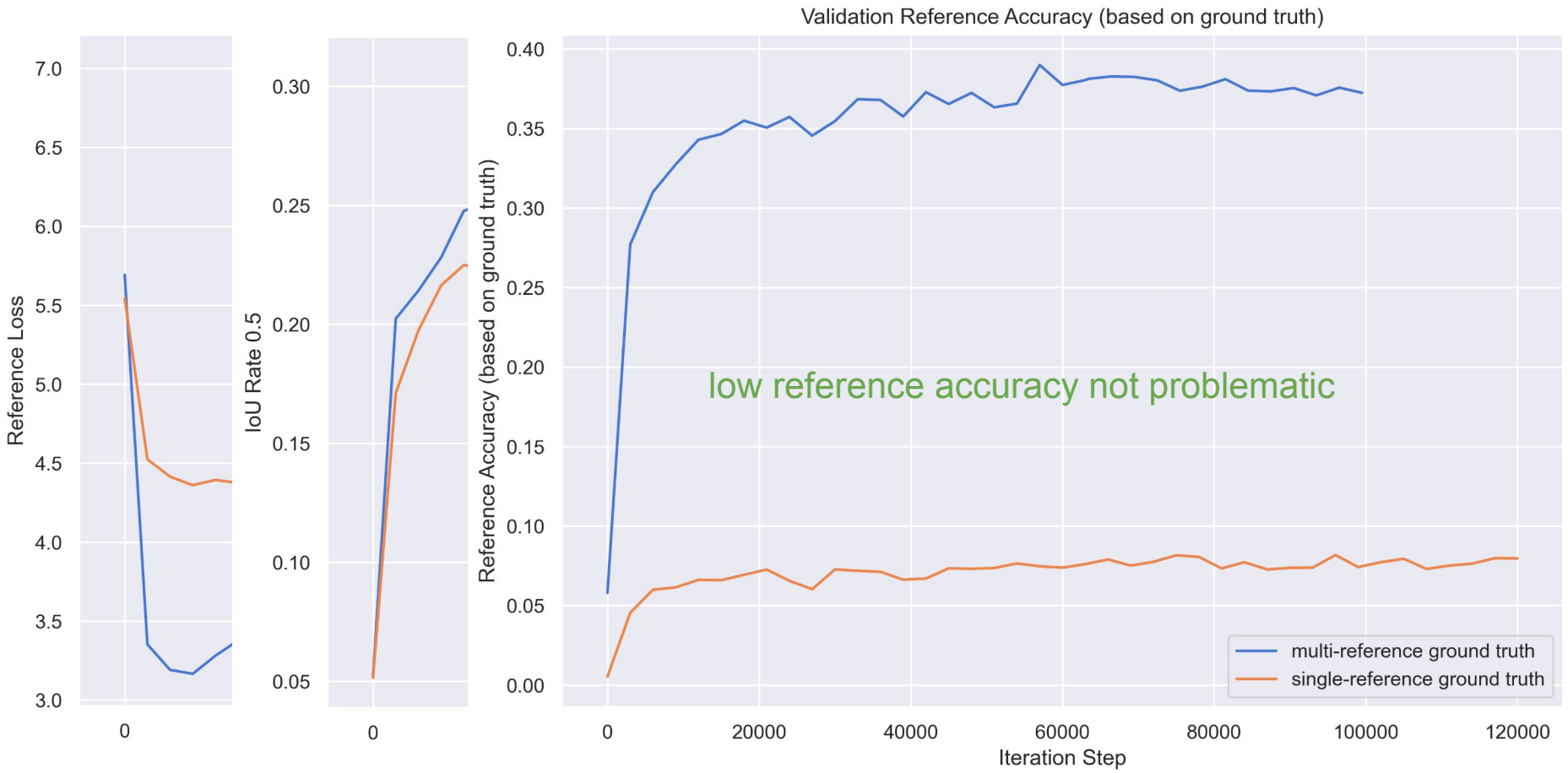
ref_{pred} = [0.91, 0.01, 0.99, 0.5, …, 0.1]

Multi Reference Experiments

One scene - Train

Multi Reference Experiments

All scenes - Validation



Multi Reference Conclusion

Single-reference task harder → regularizes

First attempts to regularize multi-reference unsuccessful

What we did not try:

- Different threshold
- Additional input features

Continue with single-reference

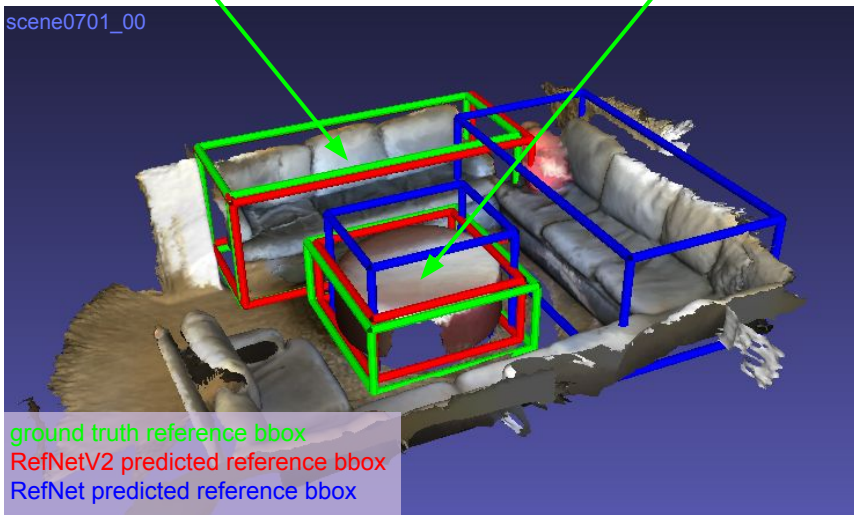
Results

	Unique Acc@0.25IoU	Unique Acc@0.50IoU	Multiple Acc@0.25IoU	Multiple Acc@0.50IoU	Overall Acc@0.25IoU	Overall Acc@0.50IoU
RefNet (XYZ, height)	63.98	43.57	29.28	18.99	36.01	23.76
RefNetV2 (XYZ) L6 pre-trained detector	71.04	57.12	22.22	17.35	31.70	25.06
RefNetV2 (XYZ),L6 detector with frozen layers	72.75 (up to pred. heads) **76.01** **(only backbone)**	58.25 (up to pred. heads) 57.21 (only backbone)	26.57 (up to pred. heads) 31.79 (only backbone)	19.75 (up to pred. heads) 22.04 (only backbone)	35.53 (up to pred. heads) 40.04 (only backbone)	27.22 (up to pred. heads) 28.86 (only backbone)
RefNetV2 (XYZ) L6 fine tuned detector	72.15 72.22 (language cls) 74.47 (hparam tuning)	54.48 53.64 (language cls) **57.78** **(hparam tuning)**	34.37 33.74 (language cls) **34.39** **(hparam tuning)**	23.48 23.61 (language cls) **25.17** **(hparam tuning)**	41.70 41.20 (language cls) **42.16** **(hparam tuning)**	29.50 29.44 (language cls) **31.50** **(hparam tuning)**
RefNetV2 (XYZ) L12 fine tuned detector	-	-	-	-	-	-
RefNet (XYZ, height, multiview, normals)	78.22	52.38	33.61	20.77	42.27	26.90
RefNetV2 (XYZ, height, rgb, normals) L6	-	-	-	-	-	-

Grounding Visualization

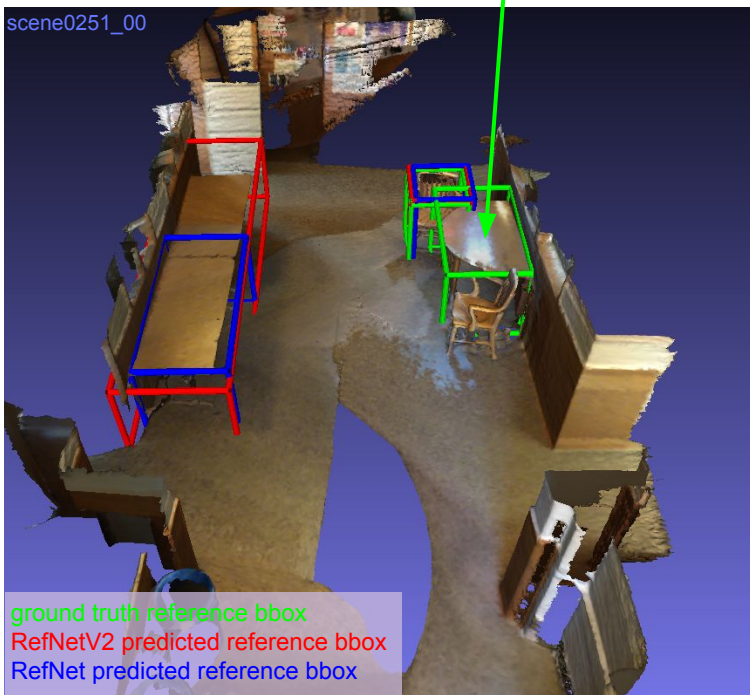
“this is a gray couch.
it is by the sectional couch.”

“this is a round table.
it is in between two couches.”



IoU Scores:	table	couch
RefNetV2	70.34	80.94
RefNet	36.00	4.73

“standing **semi-circular** wooden object
with flat surface for holding objects.
object can be found in center of image.”



IoU Scores:	chair	table
RefNetV2	85.69	0.00
RefNet	75.91	0.00

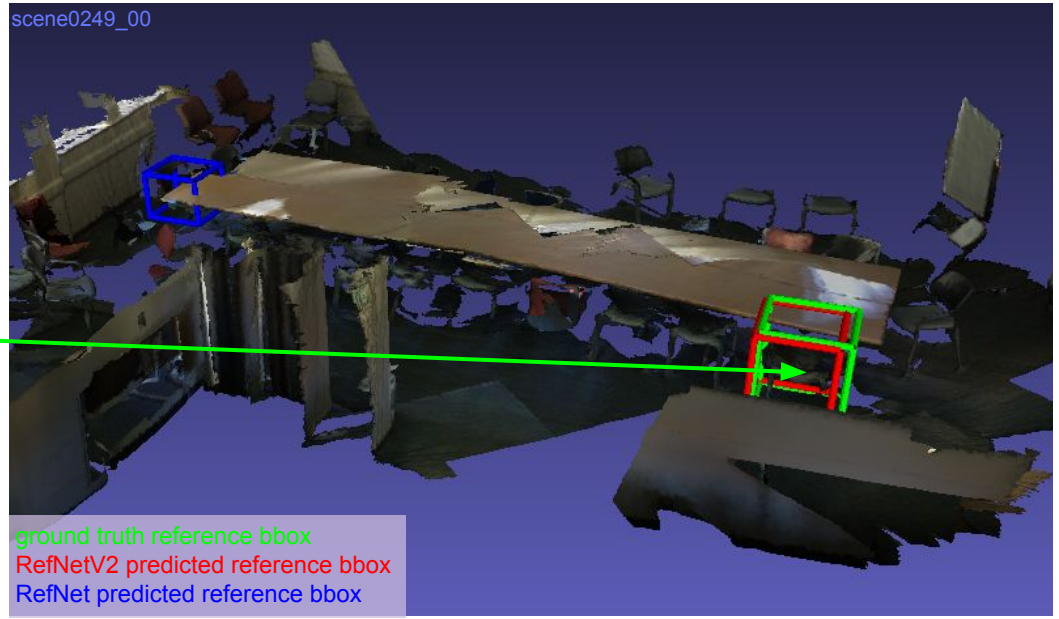
Grounding Visualization

desc1: "the black chair, it was placed near a brown table, positioned right on top of the table, **near the wall**. to the right was a white table."

desc2: "the chair is the northwestern-most one on the right side of the table. the chair is gray and has four legs."

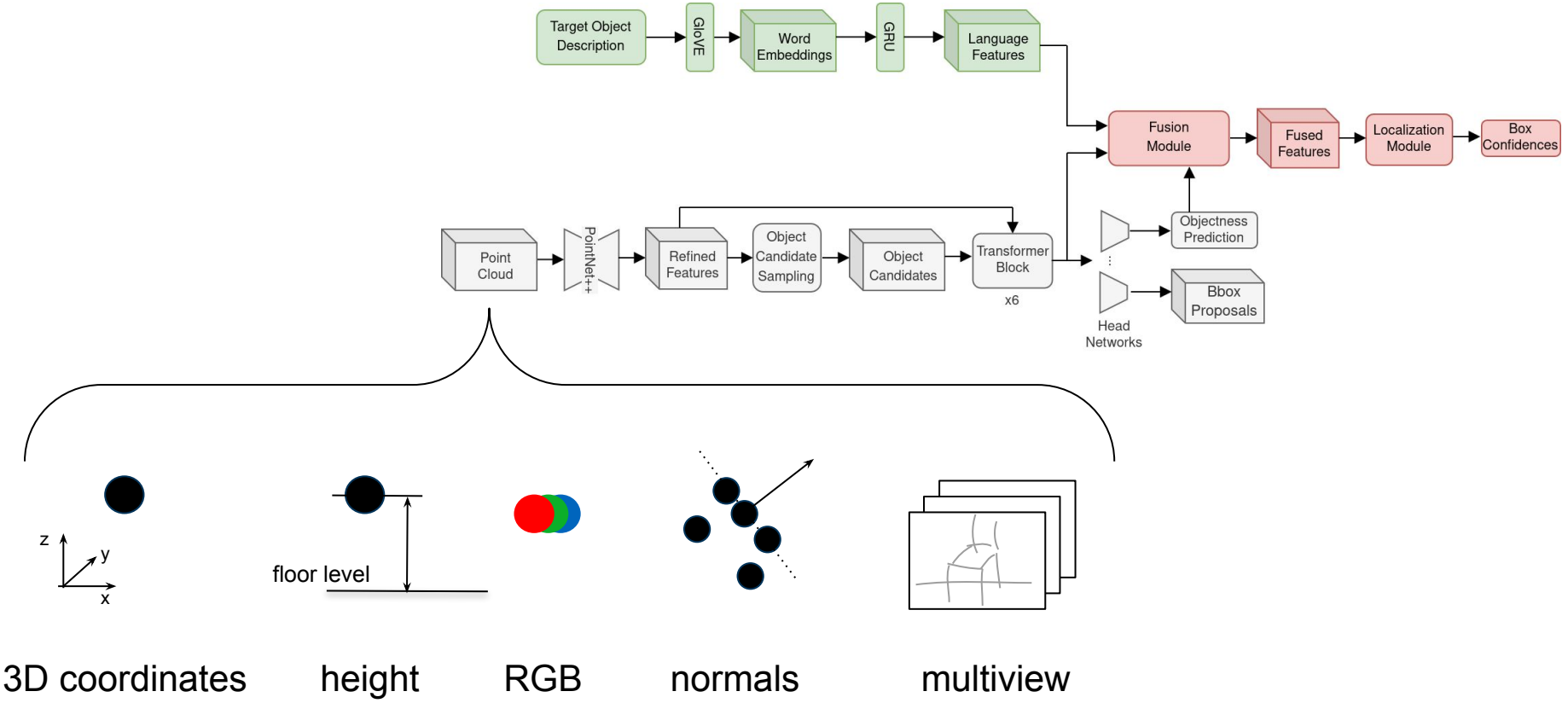
desc3: "this is a black office chair. it's on the corner of the table **closest to the whiteboard**."

desc4: "the chair is the northeastern-most one next to the table. the chair is gray and has four legs."



IoU Scores:	chair - desc1	chair - desc2	chair - desc3	chair - desc4
RefNetV2	76.22	0.00	75.52	0.00
RefNet	0.00	0.00	0.00	0.00

More Input Features - Available Features



More Input Features - ScanRefer Dataset

ScanRefer dataset is a subset of ScanNet

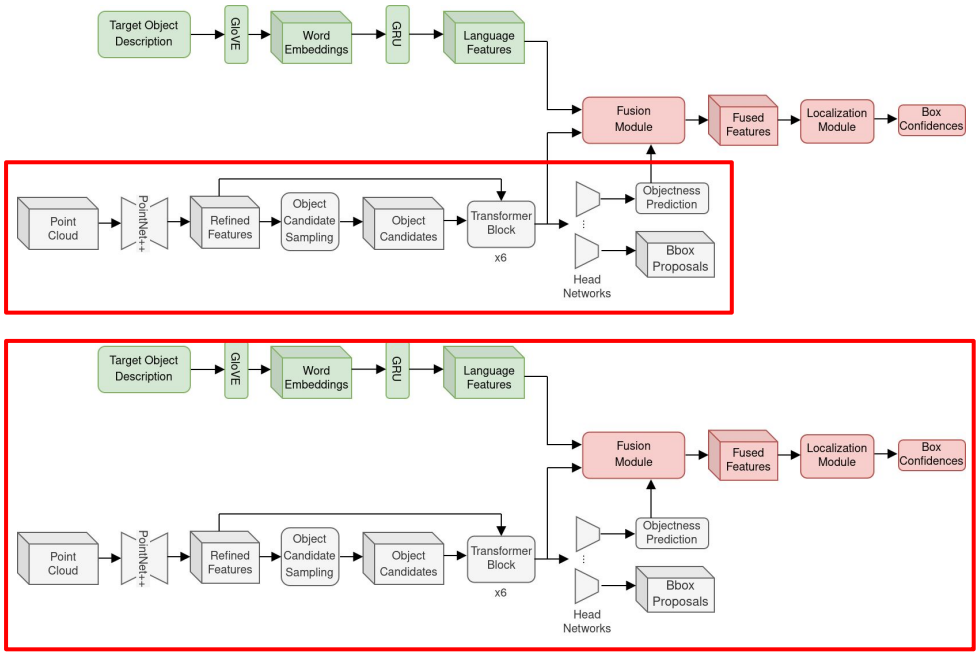
Reference performance with and without pre-trained detector (on ScanNet):

Acc@0.25IoU: **41.70** vs. 34.96

Acc@0.50IoU: **29.50** vs. 18.85

1) Pre-train detector
(on ScanNet)

2) Train RefNetV2
(on ScanRefer)



Open Challenges And Next Steps

Challenges:

- Extensive hyperparameter search almost impossible
- ScanRefer dataset not as many scans as ScanNet
- Training and evaluating different loss functions and configurations very time consuming

Next steps:

- Incorporate more input features
- Prepare poster and finish final report

References

[1] Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)

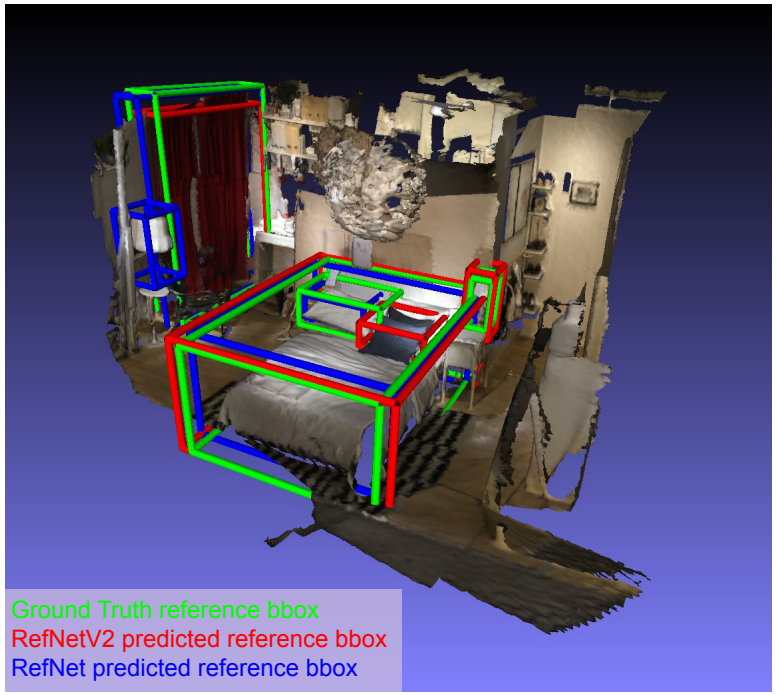
[2] Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.:Group-Free 3D Object Detection via Transformers. arXiv preprint arXiv:2104.00678 (2021)

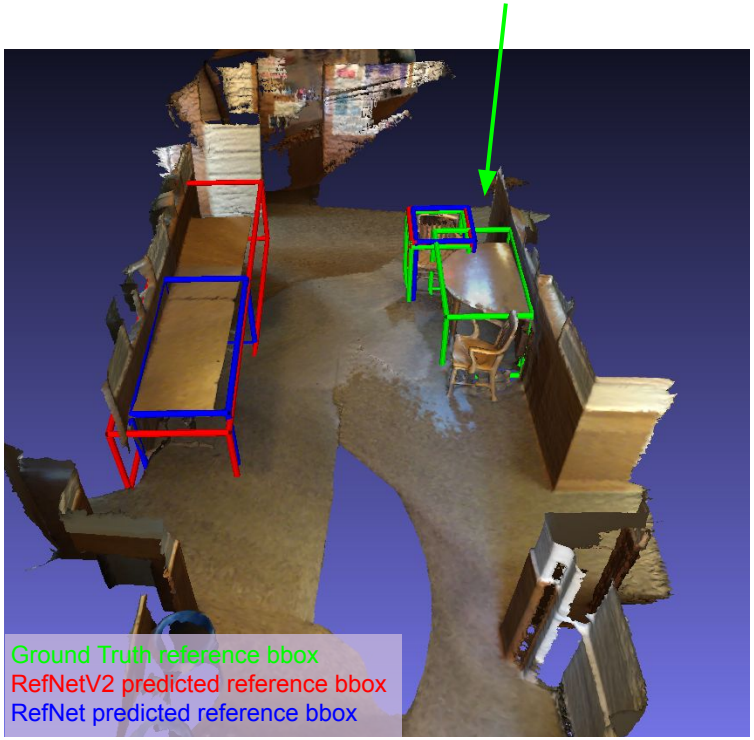
Thank you for your attention!

Appendix

Grounding Visualization

"standing semi-circular wooden object with flat surface for holding objects. object can be found in center of image."

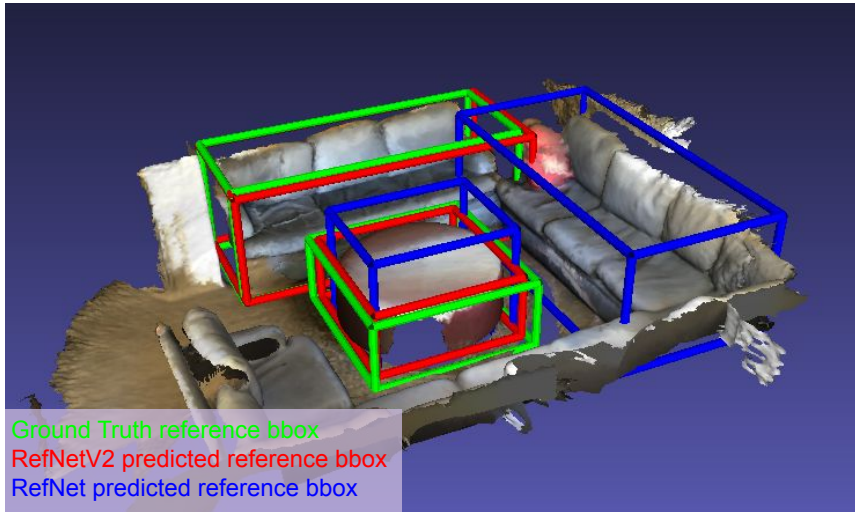


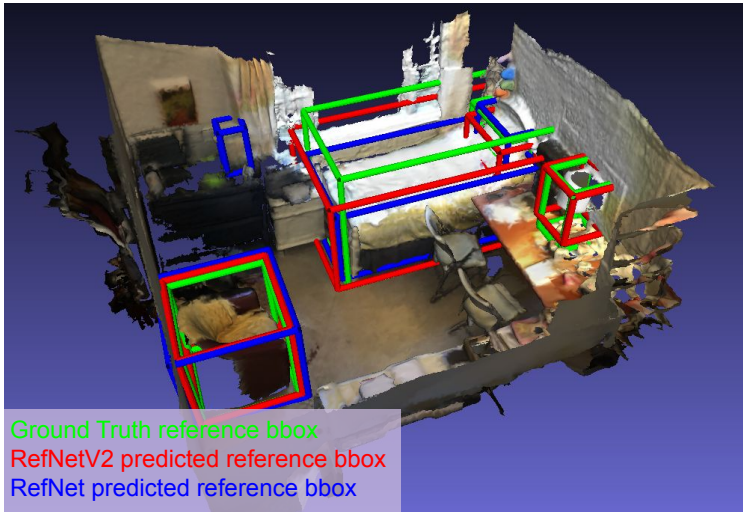


scene0246_00 **Reference IoU**	curtain	pillow	lamp	bed
RefNetV2	73.57	3.22	46.70	84.25
RefNet	67.27	40.75	0.00	74.65

scene0251_00 **Reference IoU**	chair	table
RefNetV2	85.69	0.00
RefNet	75.91	0.00

Grounding Visualization



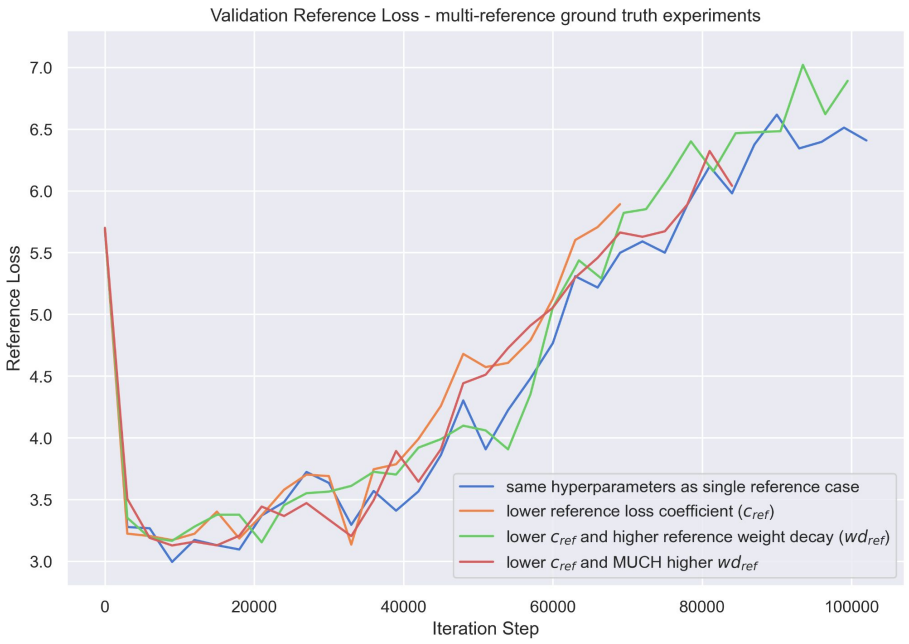


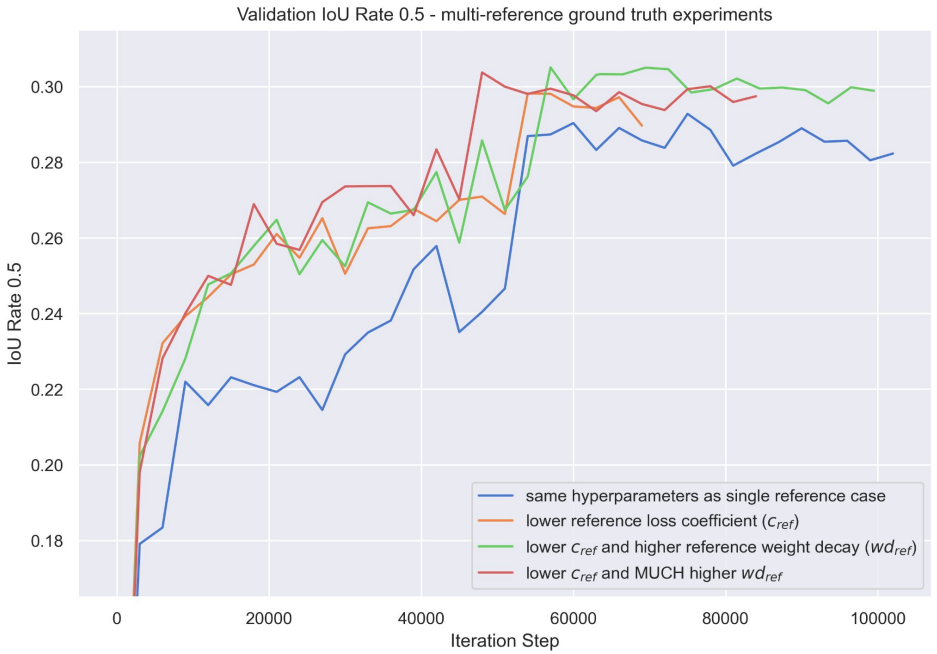
scene0701_00 **Reference IoU**	table	couch
RefNetV2	70.34	80.94
RefNet	36.00	4.73

scene0144_00 **Reference IoU**	bed	pillow	armchair	lamp
RefNetV2	62.80	40.26	80.74	59.66
RefNet	49.64	58.33	69.28	0.00

Multi Reference

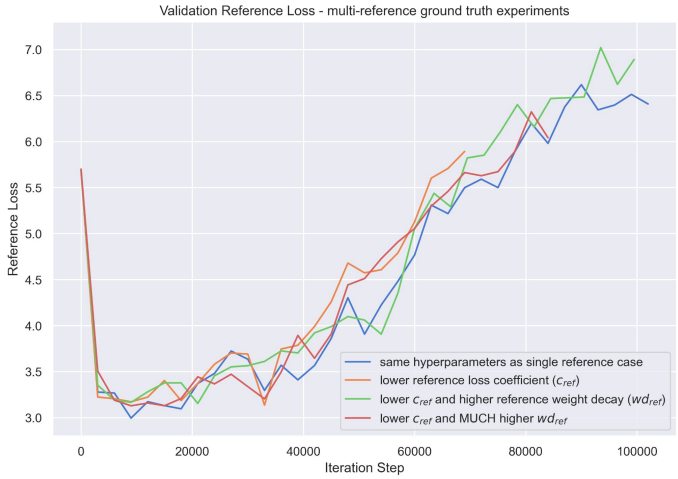
Multi reference ground truth experiments: All scenes - Validation

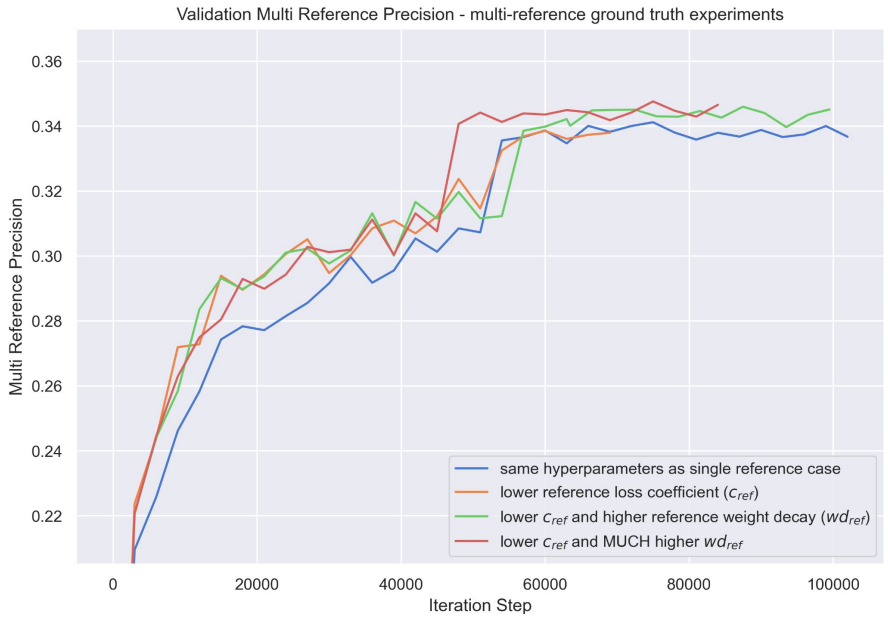


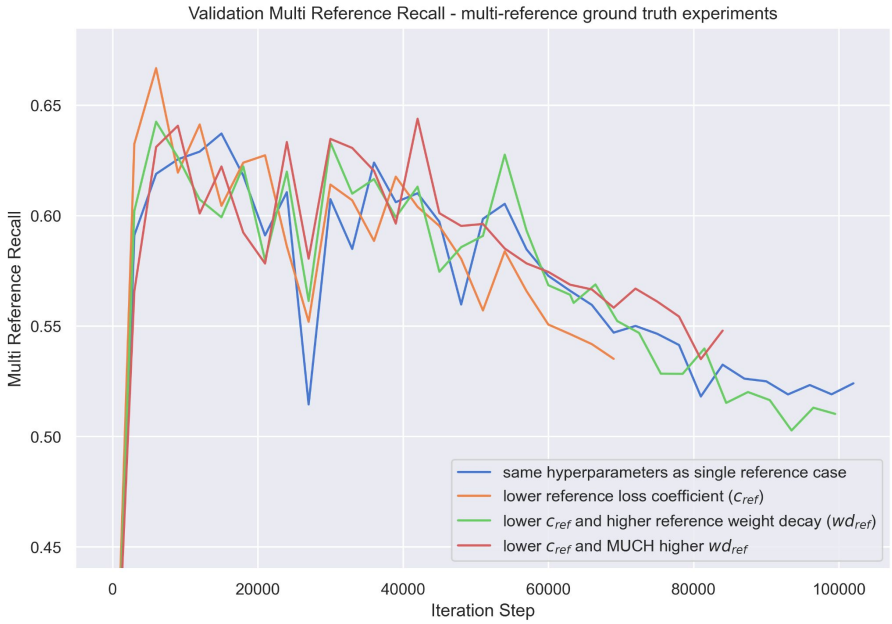


Appendix Multi-Ref

Multi reference overfitting:

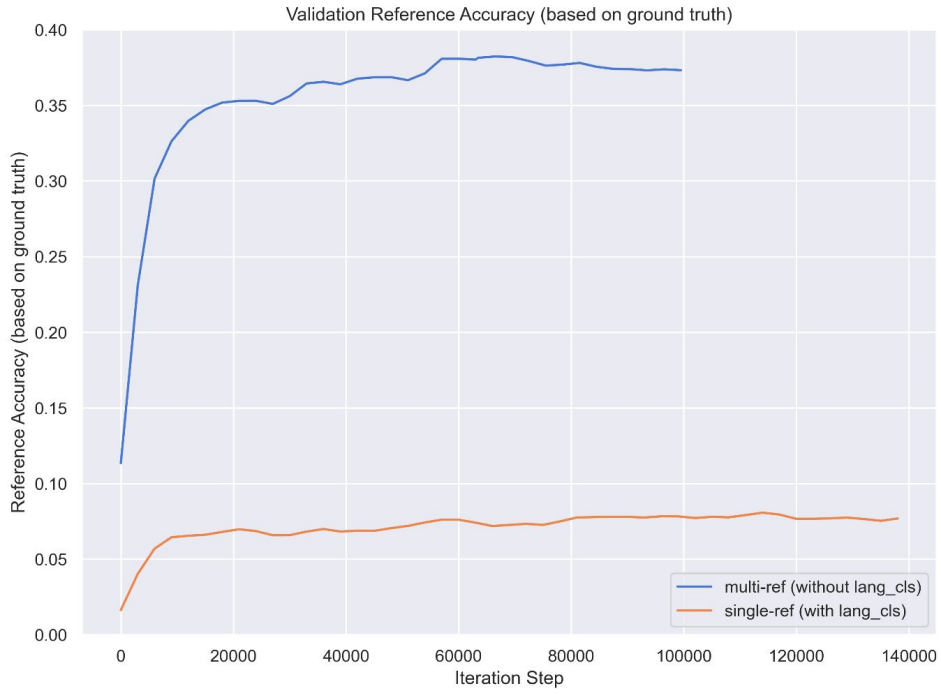


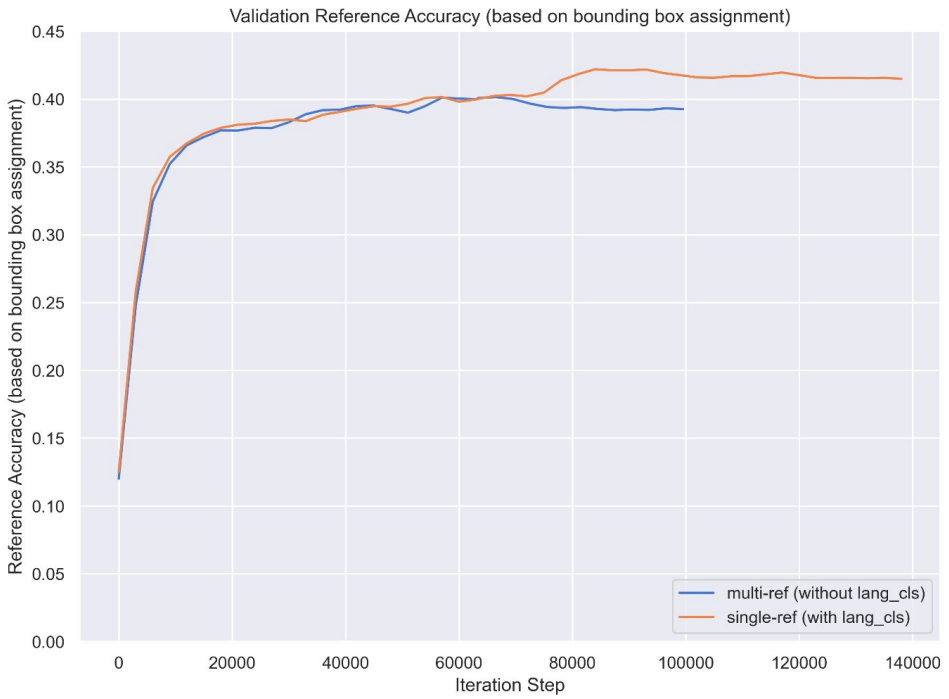




Appendix Reference Accuracy

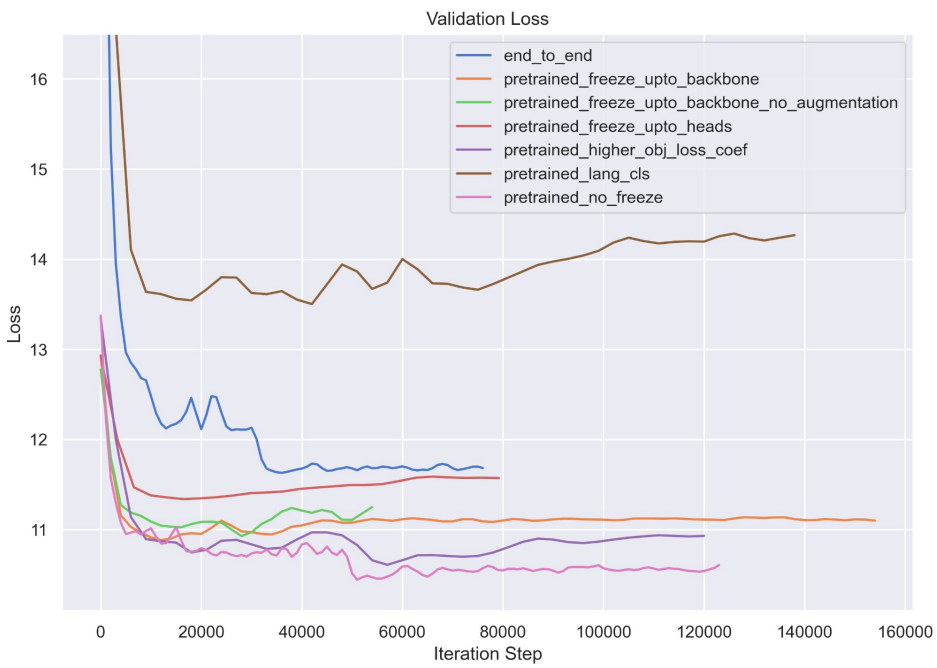
Reference accuracy based on object assignment vs based on ground truth:

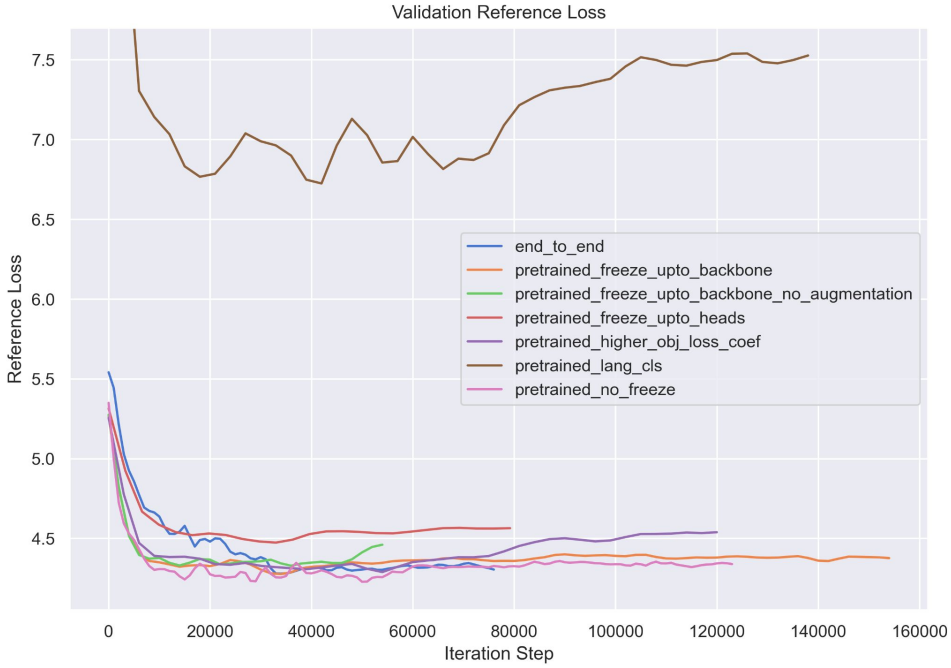




Appendix Experiments

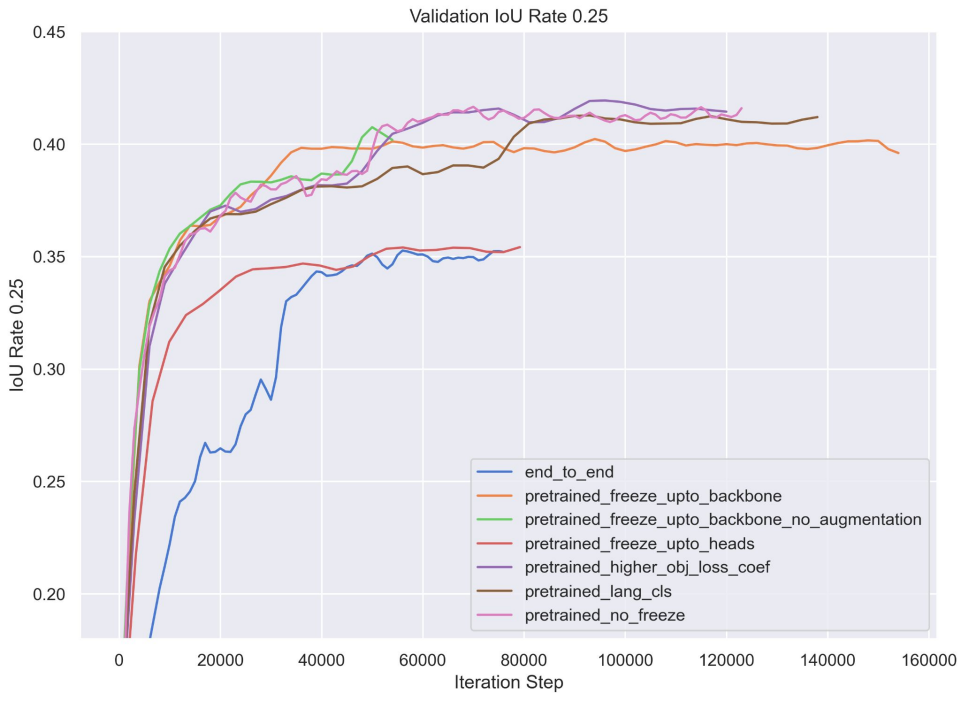
Loss curves:

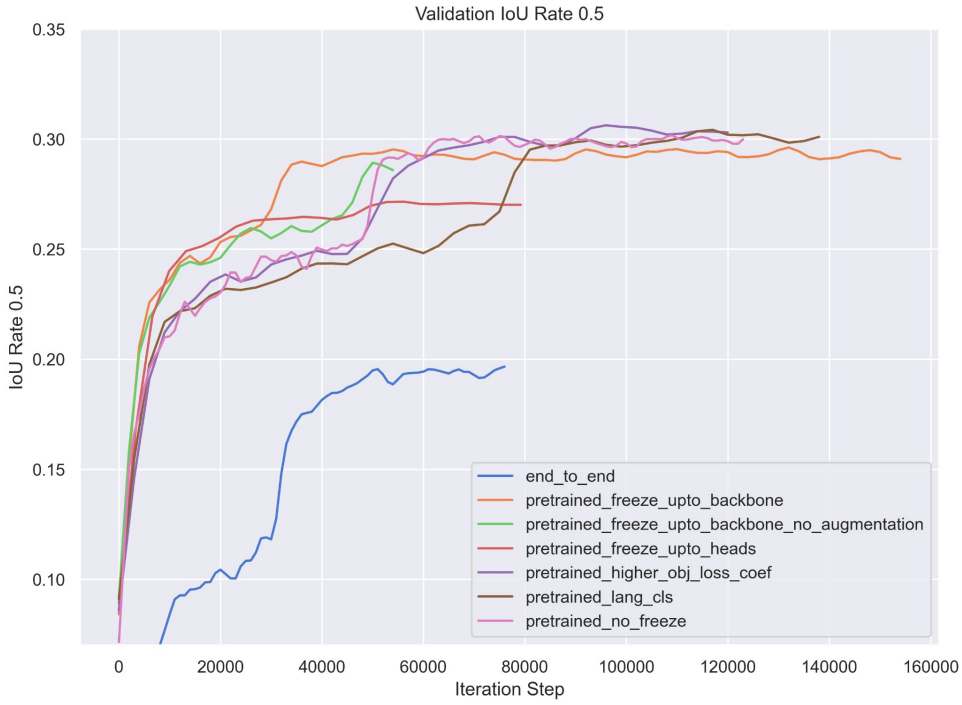


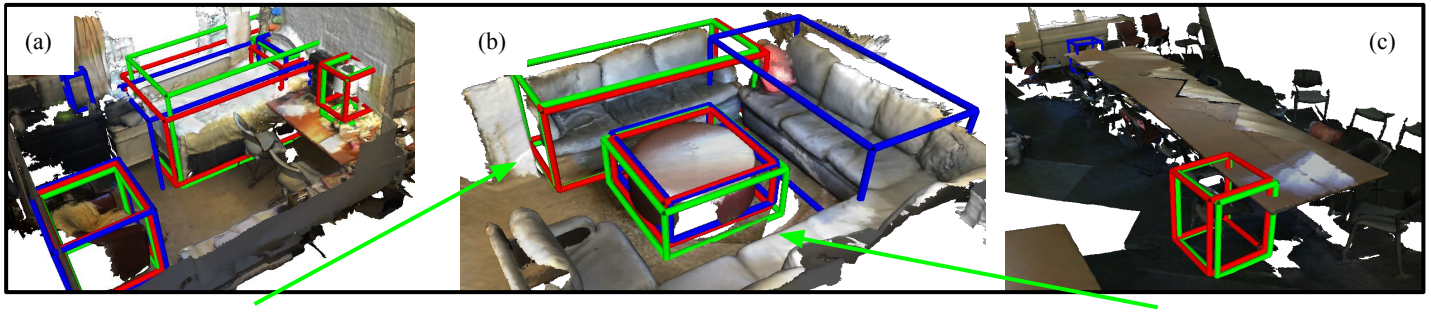


Appendix Experiments

IoU Rates:







“this is a gray couch. it is by the sectional couch.”

“this is a round table. it is in between two couches.”