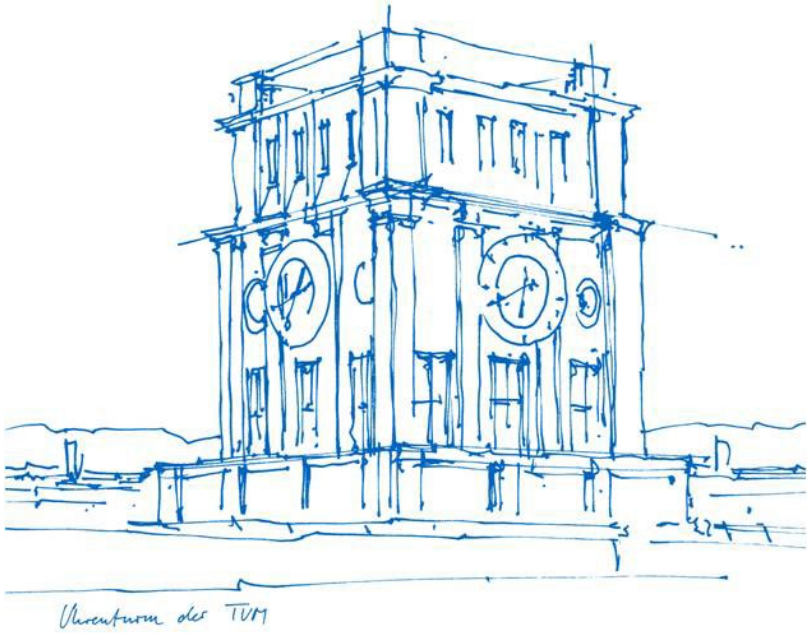
Dense Captioning for 3D Scenes

Speaker: Yunxiang Lu and Jiachen Lu

Supervisor: Dave Zhenyu Chen

Advanced Deep Learning for Computer Vision

Garching b. München, 31. Mai 2023



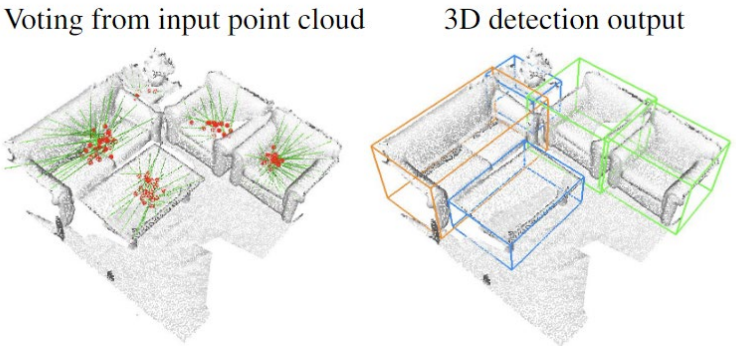
Motivation

Input: 3D Point Clouds → Bounding Boxes and Descriptions

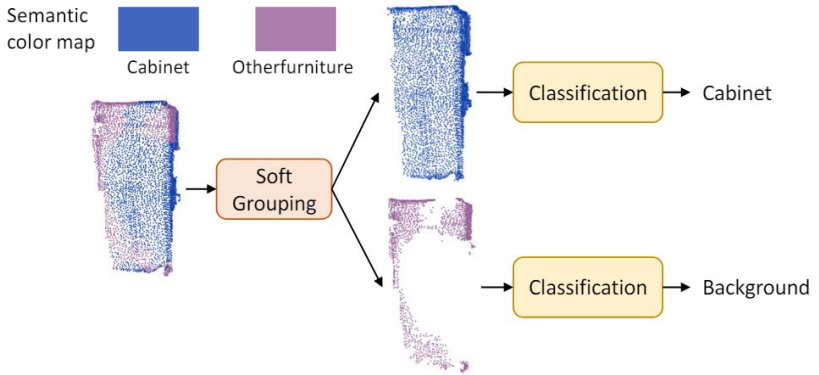
	Object Detection
Baseline (Scan2Cap)	VoteNet
Ours	SoftGroup

Better Object Detection
↓
Better Object features
↓
Better Object Captions

Motivation



VoteNet mechanism



SoftGroup mechanism

	mAP@0.25IoU	mAP@0.5IoU
VoteNet	58.6	33.5
SoftGroup	**71.6**	**59.4**

SoftGroup-based network architecture

$N_p \times 6$

point clouds

Voxelization

U-Net

Devoxelization

Semantic branch

Offset branch

Soft Grouping

Bottom-up grouping

Instance proposals

Per-instance feature extractor

Tiny U-Net

Classification

Segmentation

Mask scoring

Final Instances (Bounding boxes)

GRU

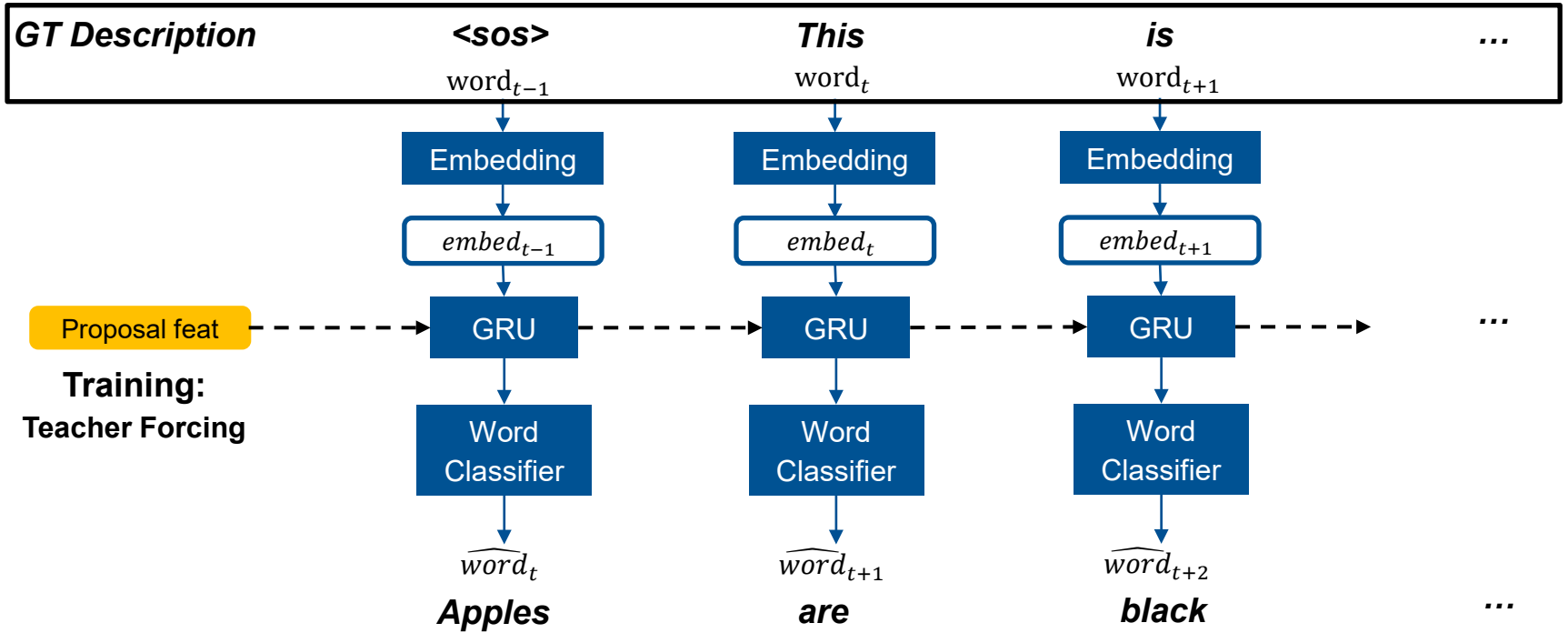
Object Descriptions

Object Features *Mx32*

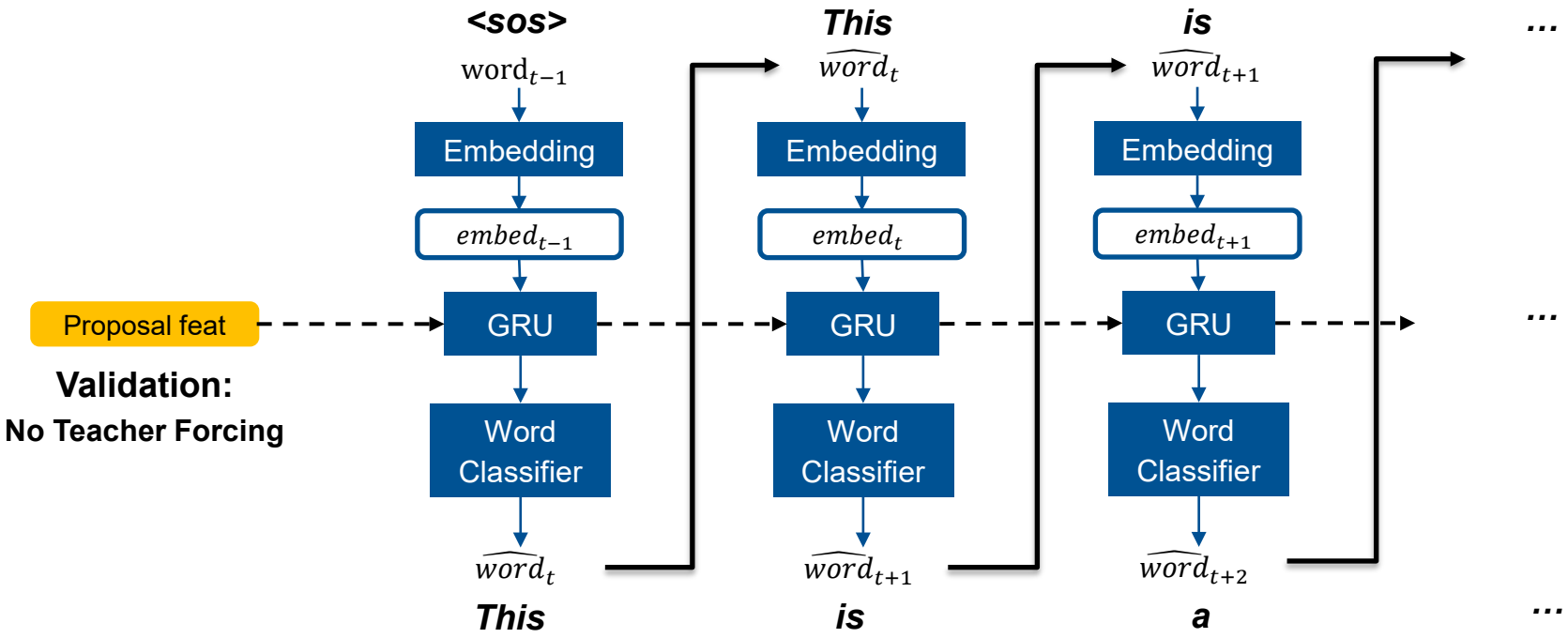
SoftGroup-based Backbone

Top-down refinement

SoftGroup-based network architecture

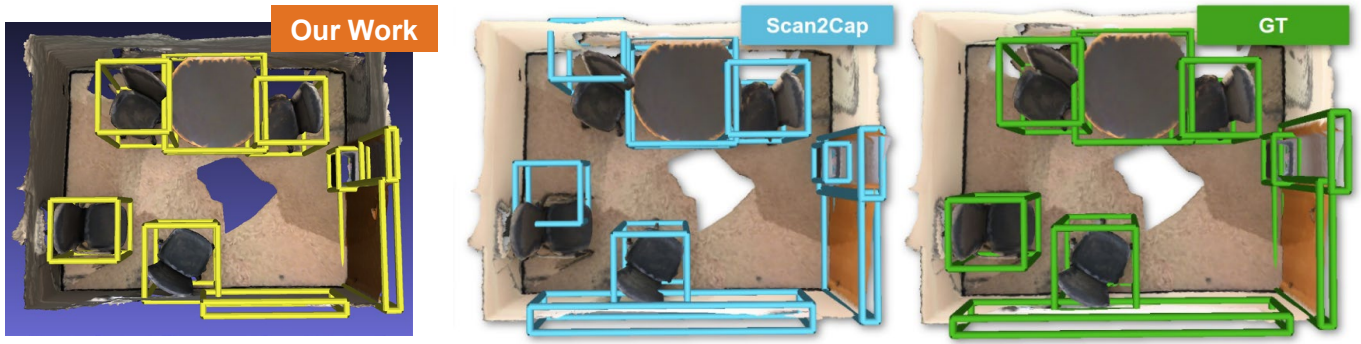


SoftGroup-based network architecture



Training and Inference results

	Network Architecture	CIDEr @0.5IoU	BLEU-4 @0.5IoU	METEOR @0.5IoU	ROUGE @0.5IoU	Box mAP @0.5IoU
Scan2Cap	VoteNet+GRU	34.31	21.42	20.13	41.33	32.21
Ours	SoftGroup+GRU	**34.70**	**24.46**	**22.23**	**46.95**	**45.56**

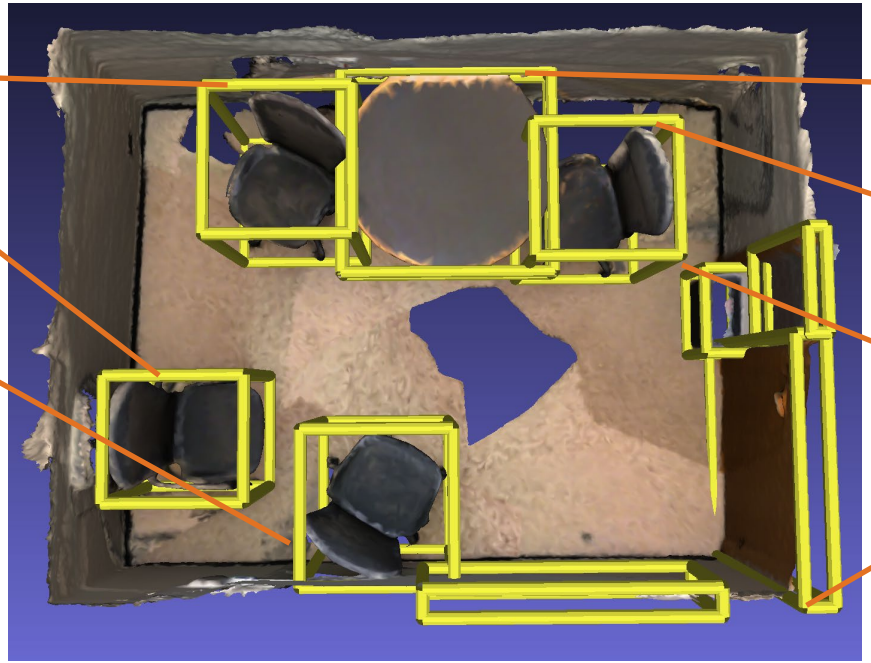


Training and Inference results

This is a black office chair. It is in the corner of the room.

This is a black office chair. It is in the corner of the room.

This is a black office chair. It is in the corner of the room.



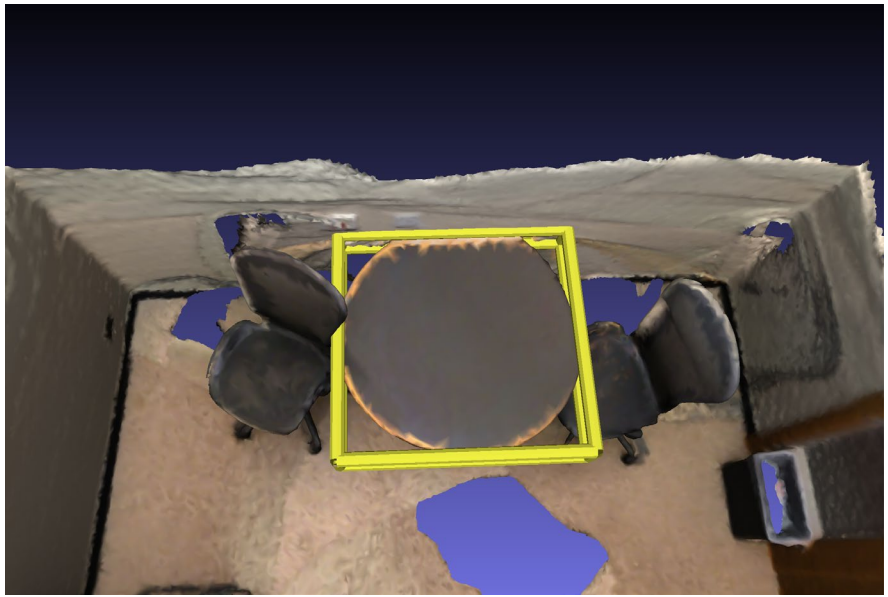
This is a brown table. It is in the center of the room.

This is a black office chair. It is in the corner of the room.

The trash can is blue. It is located to the right of the trash can.

The door is white. It is located to the right of the trash can

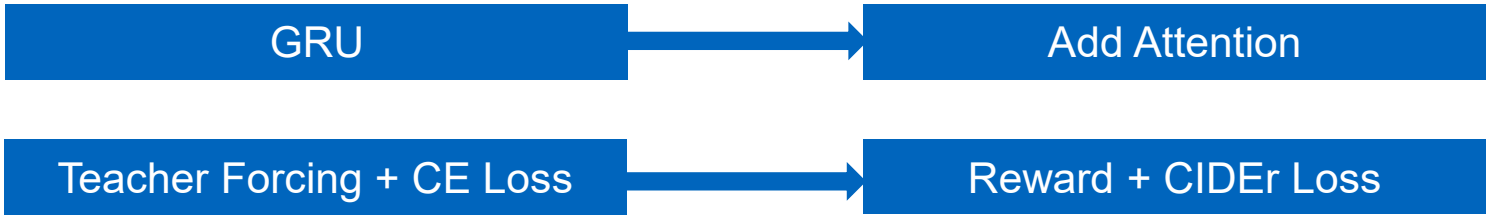
Training and Inference results



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

Inference:
This is a **brown table**. It is **in the center of** the **room**

Challenges and Future works



Thank you for your attention!

Any Question?

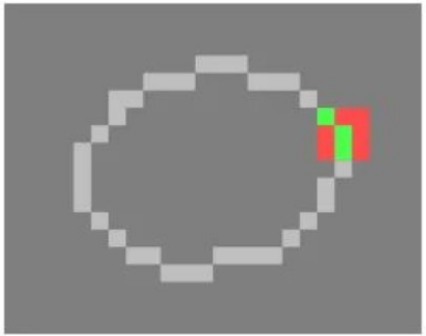
Additional materials

Sparse Convolution
Loss Formula
…

Additional materials

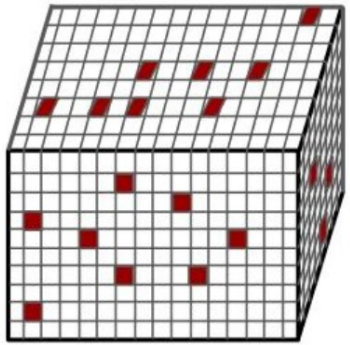
Sparse Convolution

- Regular: calculate output when kernel covers an active input
- Submanifold: calculate output only when the kernel center covers an active input



Reason:

- In 3D Space, many voxels are empty, so the point clouds data are always sparse.
- Using Sparse Convolution can help us calculate the feature more efficiently



Additional materials

Loss Formula

$$L_{\text{semantic}} = \frac{1}{N}\sum_{i=1}^{N}\text{CE}(\boldsymbol{s}_i, s_i^*),$$

$$L_{\text{offset}} = \frac{1}{\sum_{i=1}^{N}\mathbb{1}_{\{\boldsymbol{p}_i\}}}\sum_{i=1}^{N}\mathbb{1}_{\{\boldsymbol{p}_i\}}\|\boldsymbol{o}_i - \boldsymbol{o}_i^*\|_1,$$

$$L_{\text{class}} = \frac{1}{K}\sum_{k=1}^{K}\text{CE}(\boldsymbol{c}_k, c_k^*),$$

$$L_{\text{mask}} = \frac{1}{\sum_{k=1}^{K}\mathbb{1}_{\{\boldsymbol{m}_k\}}}\sum_{k=1}^{K}\mathbb{1}_{\{\boldsymbol{m}_k\}}\text{BCE}(\boldsymbol{m}_k, \boldsymbol{m}_k^*),$$

$$L_{\text{mask\_score}} = \frac{1}{\sum_{k=1}^{K}\mathbb{1}_{\{\boldsymbol{e}_k\}}}\sum_{k=1}^{K}\mathbb{1}_{\{\boldsymbol{e}_k\}}\|\boldsymbol{e}_k - \boldsymbol{e}_k^*\|_2.$$

GT:

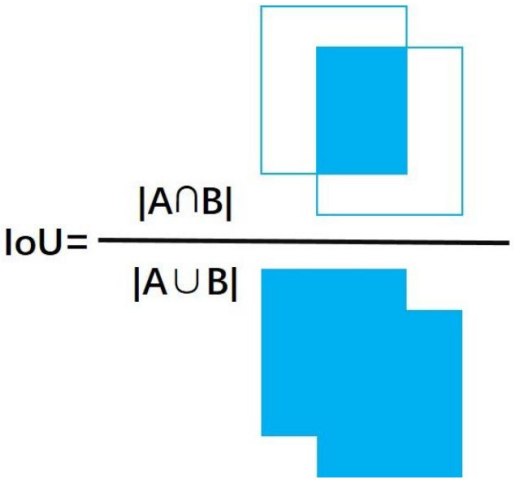
- Semantic: GT semantic class of the point
- Offset: GT offset vector that shifts the point to corresponding instance center

- Classification: class of the GT instance with highest IoU
- Mask: the mask of the assigned GT instance
- Mask score: IoU between the predicted mask and the GT

We treat all instance proposals having IoU with a ground-truth instance higher than 50% as the positive samples and the rest as negatives.

Additional materials

IoU: Intersection of Union



Program Results (NLP metrics & Box mAP@0.5IoU)

```
loading corpus...
generating descriptions...
computing scores...
BLEU-1,2,3,4 are: [0.5248459993766885, 0.41716240214357864, 0.31902357245013113, 0.24464104223299268]
CIDEr is: 0.3470318672484949
BLEU-4 is: 0.24464104223299268
METEOR is: 0.22290531565740027
ROUGE is: 0.4694894696899399
```

```
Processing scene0704_00
Evaluating...
mAP: 0.4555707314393957
```

Description Results (>0.5 bounding box iou)


```
"scene0427_00|5|chair": [
    "sos this is a black chair . it is in the corner of the room . eos"
],
"scene0427_00|8|door": [
    "sos the door is white . it is located to the right of the trash can . eos"
],
"scene0427_00|0|trash_can": [
    "sos the trash can is blue . it is located to the right of the trash can . eos"
],
"scene0427_00|7|chair": [
    "sos this is a black chair . it is in the corner of the room . eos"
],
"scene0427_00|9|table": [
    "sos this is a brown table . it is in the center of the room . eos"
],
"scene0427_00|4|chair": [
    "sos this is a black chair . it is in the corner of the room . eos"
],
"scene0427_00|6|chair": [
    "sos this is a black chair . it is in the corner of the room . eos"
],
"scene0427_00|2|window": [
    "sos eos"
```