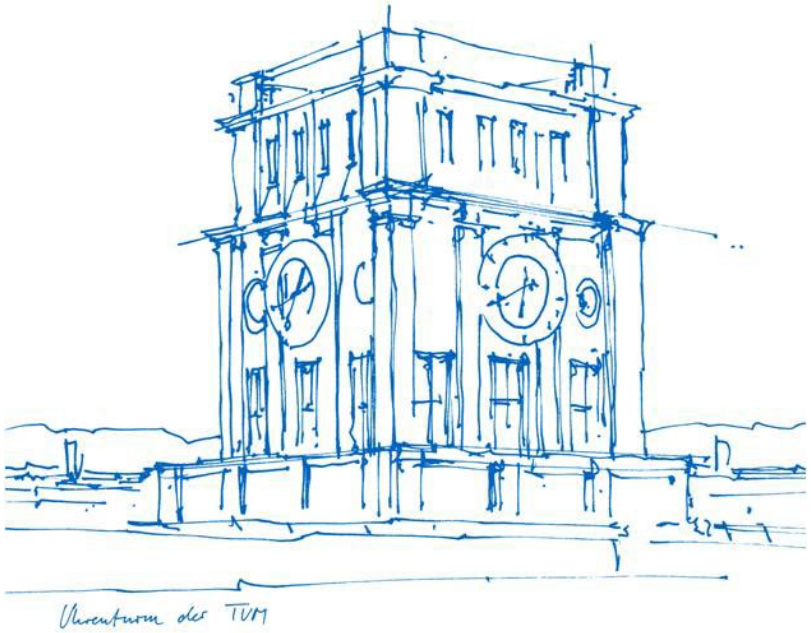
Dense Captioning for 3D Scenes

Speaker: Yunxiang Lu and Jiachen Lu

Supervisor: Dave Zhenyu Chen

Advanced Deep Learning for Computer Vision

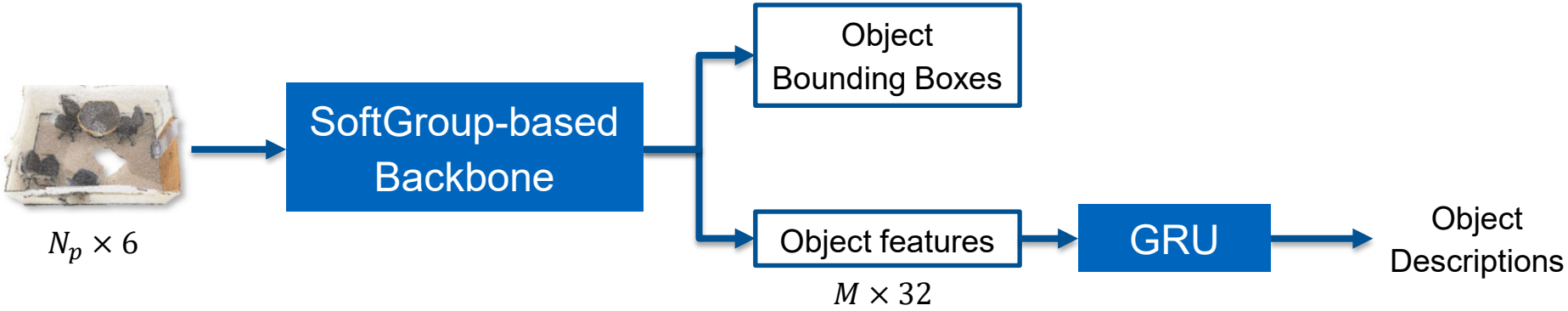
Garching b. München, 28. June 2023



Recall

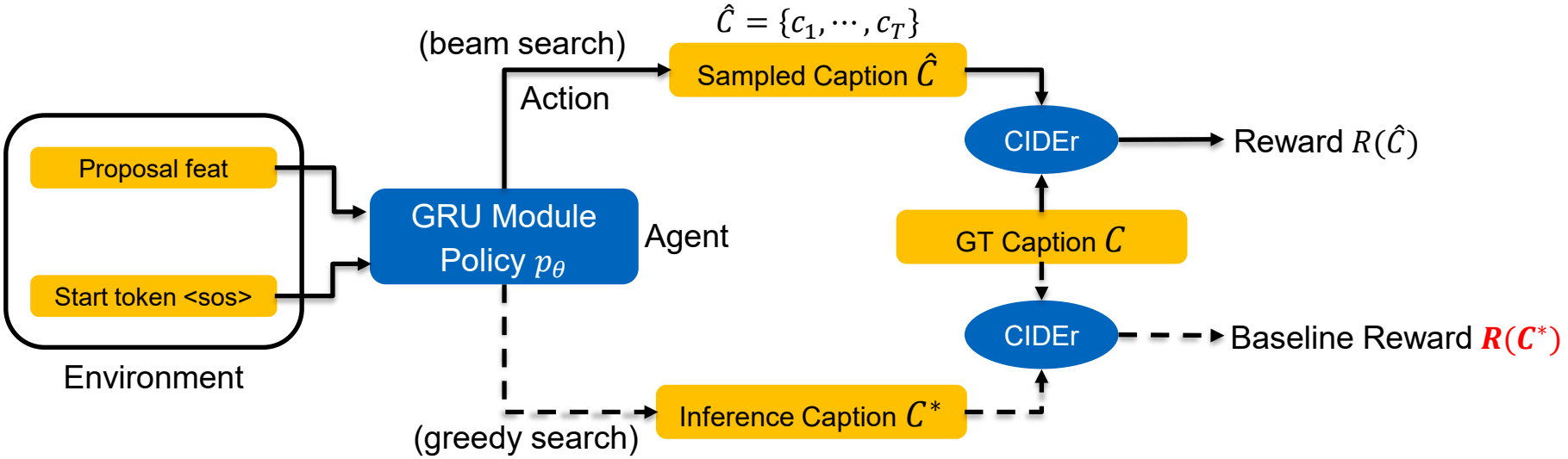

it is a black office chair next to the round table. it is against the wall.
this is a round table. it is between two black chairs. it is in the center of this room.
this is a black office chair. it is in the corner next to a black chair.
this is a black office chair. it is against the wall and facing the round table.
it is a black office chair. it is facing the round table.
this is a small trashcan. it is in front of a glass window.
it is a glass window. it is set next to the wooden door.
it is a brown door next to a window. it is wooden and closed.

Recall



	Network	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	**43.52**	**23.18**	**23.59**	**48.91**	**55.64**

Reinforcement Learning + CIDEr Loss

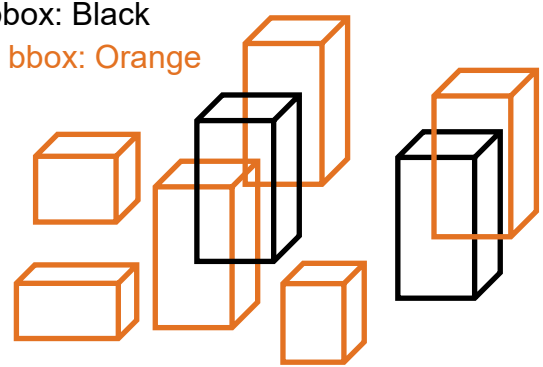


Policy Gradient: $L_{cap}(\theta) \approx -\left[R(\hat{C}) - \boldsymbol{R(C^*)}\right] \sum_{t=1}^{T} \log p(\hat{c}_t|\theta)$

Evaluation

		Detection	Captioning Recall @0.5IoU				Captioning Precision @0.5IoU				Captioning F1-Score @0.5IoU			
Model	Loss	mAP@0.5	C	B-4	M	R	C	B-4	M	R	C	B-4	M	R
SoftGroup+GRU	Cross Entropy	55.64	44.76	25.21	23.94	50.31	18.61	10.71	9.05	19.05	26.29	15.04	13.13	27.64
SoftGroup+GRU	CIDEr Loss	**55.80**	**58.24**	**29.61**	**24.32**	**51.03**	**24.05**	**12.51**	**9.18**	**19.36**	**34.05**	**17.59**	**13.34**	**28.07**

GT bbox: Black
Pred bbox: Orange



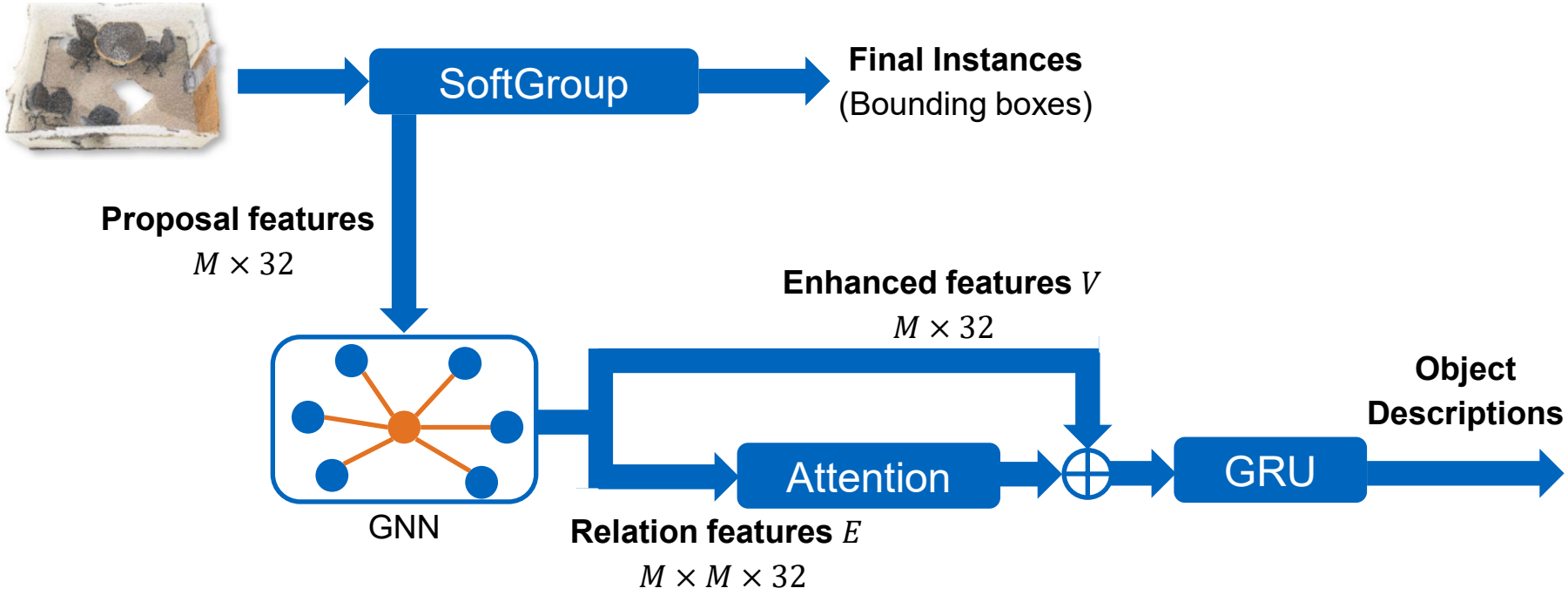
$$M^{Recall}@0.5IoU = \frac{1}{N^{GT}} \sum_{i=1}^{N^{GT}} m_i u_i$$

$$M^{Precision}@0.5IoU = \frac{1}{N^{pred}} \sum_{i=1}^{N^{pred}} m_i u_i$$

$$M@0.5IoU = \frac{2 \times M^{P}@0.5IoU \times M^{R}@0.5IoU}{M^{P}@0.5IoU + M^{R}@0.5IoU}$$

m_i: Metric score
u_i: IoU mask

Softgroup + GNN + Attention + GRU



Softgroup + GNN + CAC

SoftGroup

Final Instances
(Bounding boxes)

Proposal features
$M \times 32$

Enhanced features V
$M \times 32$

GNN

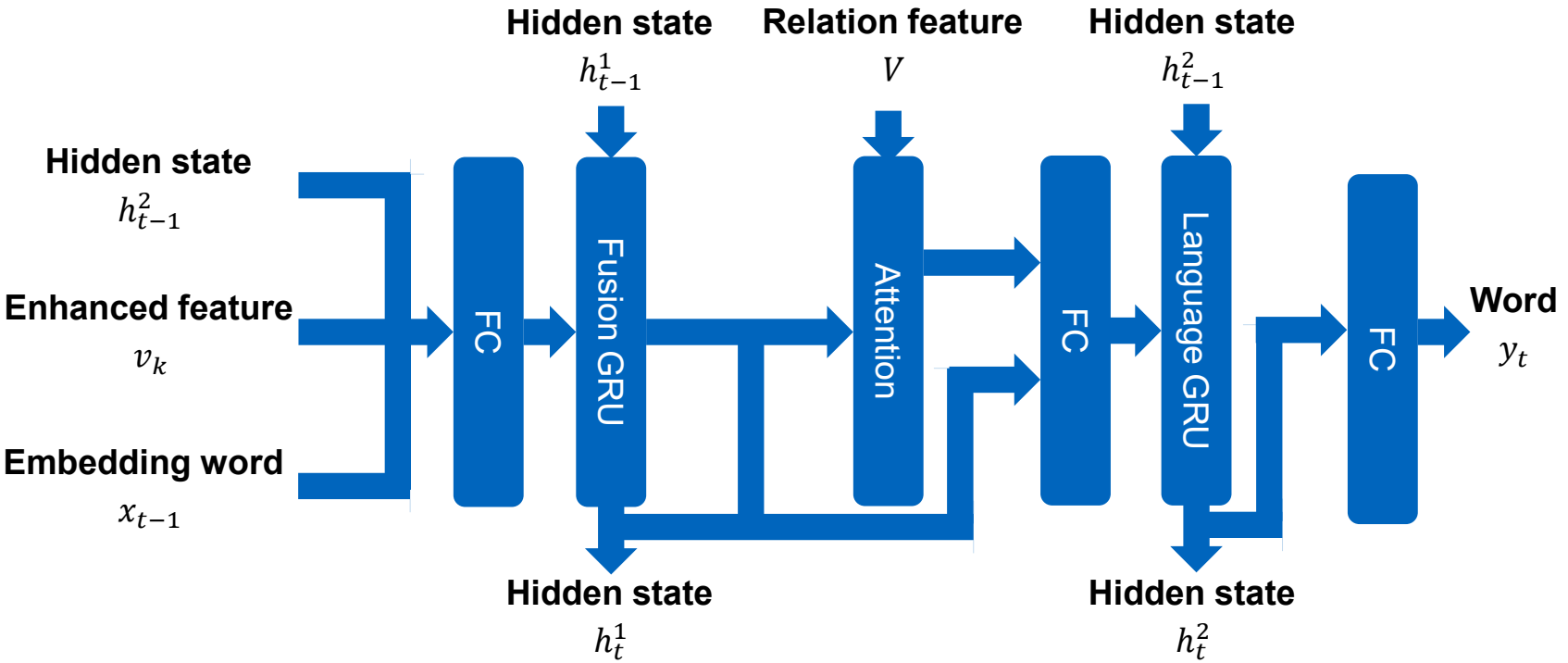
Relation features E
$M \times M \times 32$

FC + GRU
Attention
FC + GRU

CAC

Object Descriptions

Context-aware Attention Captioning (CAC)



Ablation study

		Detection	Captioning F1-Score @0.5IoU			
Loss	Network	mAP@0.5	C	B-4	M	R
CE	SoftGroup+GRU	55.64	25.69	13.74	12.96	26.88
CE	SoftGroup+RG+GRU	55.13	25.57	12.67	12.50	25.82
CE	SoftGroup+RG+Att2GRU	56.48	26.77	14.81	13.13	27.48
CE	**SoftGroup+RG+CAC**	**57.22**	**30.76**	**16.30**	**13.83**	**28.41**
CIDEr Loss	SoftGroup+GRU	55.80	33.24	17.45	13.57	28.36
CIDEr Loss	SoftGroup+RG+GRU	55.29	33.08	16.09	13.09	27.24
CIDEr Loss	SoftGroup+RG+Att2GRU	56.64	34.78	17.62	13.46	28.23
CIDEr Loss	**SoftGroup+RG+CAC**	**57.38**	**36.27**	**18.66**	**13.82**	**29.13**

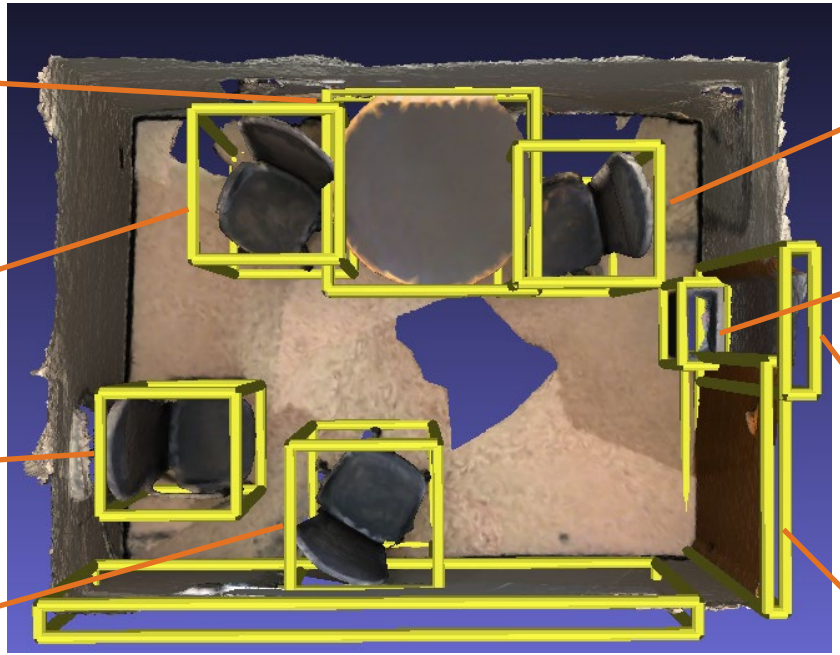
Result Visualization

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.



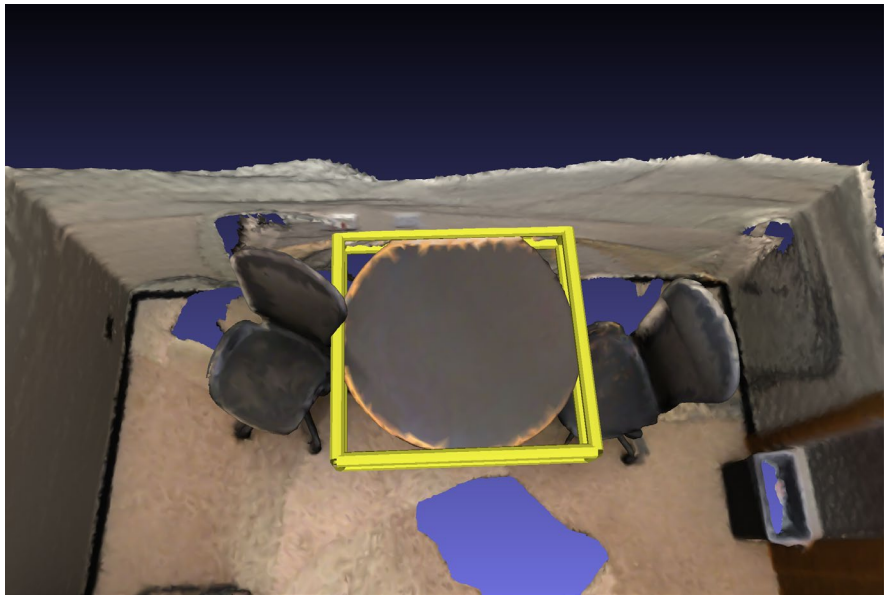
This is a black chair.
It is at a round table.

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

Result Visualization



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 Loss:

The table is a **round table**. It is **in the right of the room**.

Future works

- Final 2 weeks:
  - Conduct more ablation experiments
  - Prepare final report and presentation
  - Organize the code

- Future Possibilities:
  - GRU → Transformer
  - Listener module → Increase the discriminability of captions

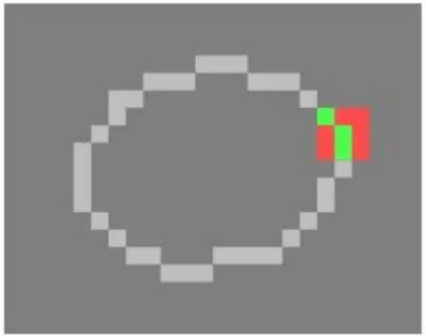
Thank you for your attention!

Any Question?

Additional materials

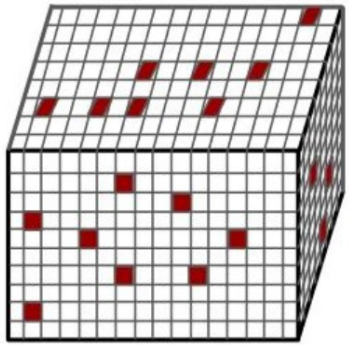
Sparse Convolution

- 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:

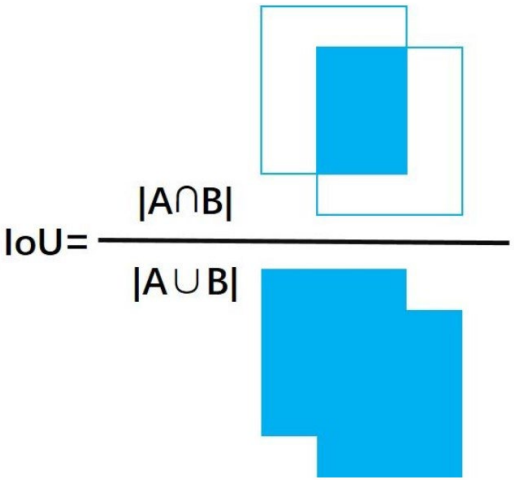
- Semantic: GT semantic class of the point
- Offset: GT offset vector that shifts the point to corresponding instance center

- 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

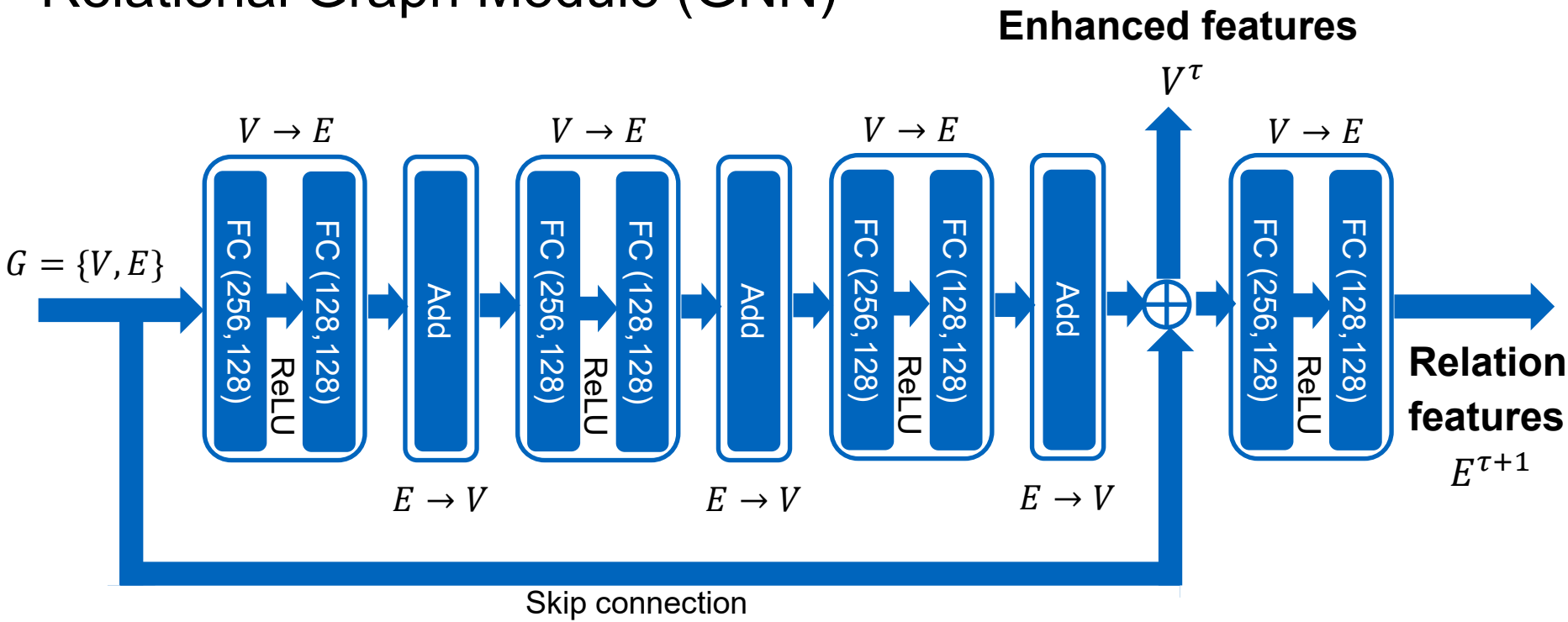
We treat all instance proposals having IoU with a ground-truth instance higher than 50% as the positive samples and the rest as negatives.

Additional materials

IoU: Intersection of Union



Relational Graph Module (GNN)



GNN

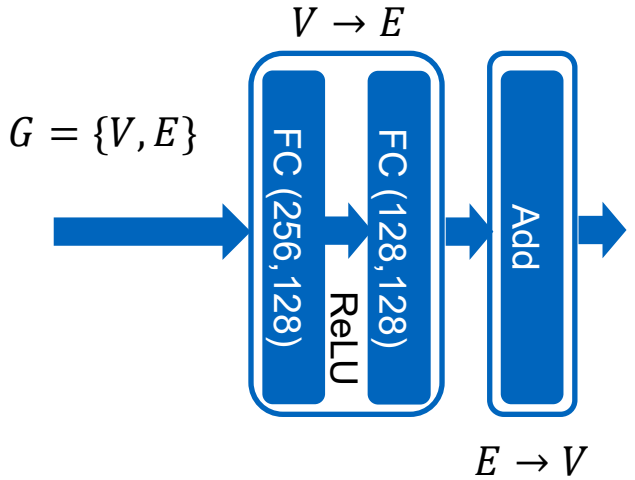
$G: Graph / V: Node / E: Edge$

Message passing

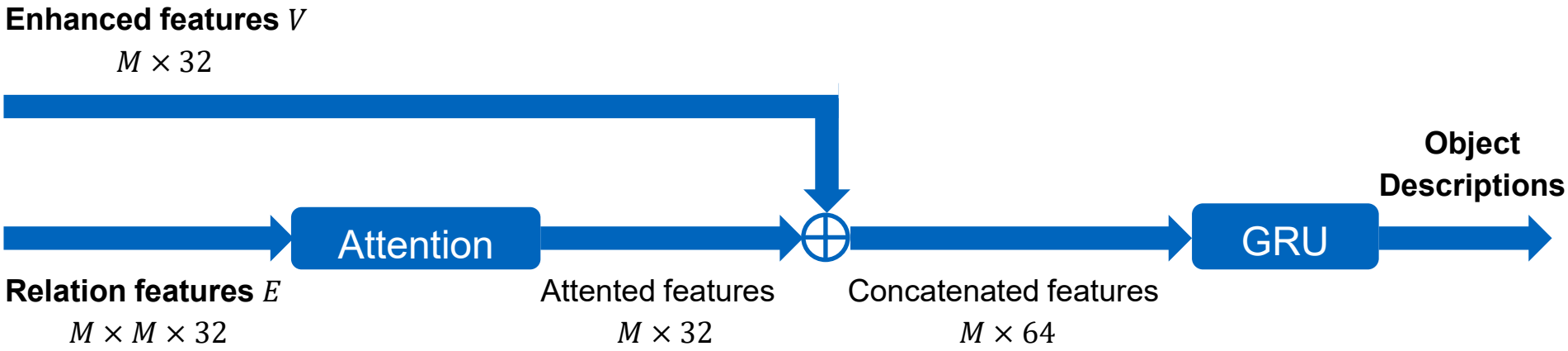
$V \rightarrow E$: $G_{i,j}^{\tau+1} = f^{\tau}([G_i^{\tau}, G_j^{\tau} - G_i^{\tau}])$

Aggregated node features

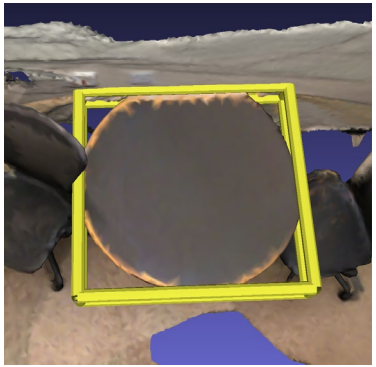
$E \rightarrow V$: $G_i^{\tau+1} = \sum_{k=1}^{K} G_{i,k}^{\tau}$



Attention+GRU



Result Visualization



GT:

"sos the round table has a black top . there are two black office chairs round the table . eos",
"sos the round table with the black top is in the corner of the room . two black office chairs around the table . eos“
"sos there is a circular table . it is in the center of the room . eos“
"sos this is a brown table that is round . it has a black circular center . eos“
"sos this is a round table top in the meeting room . there are two office chairs around it . eos"