LESS: Label-Efficient and Single-Stage Referring 3D Segmentation

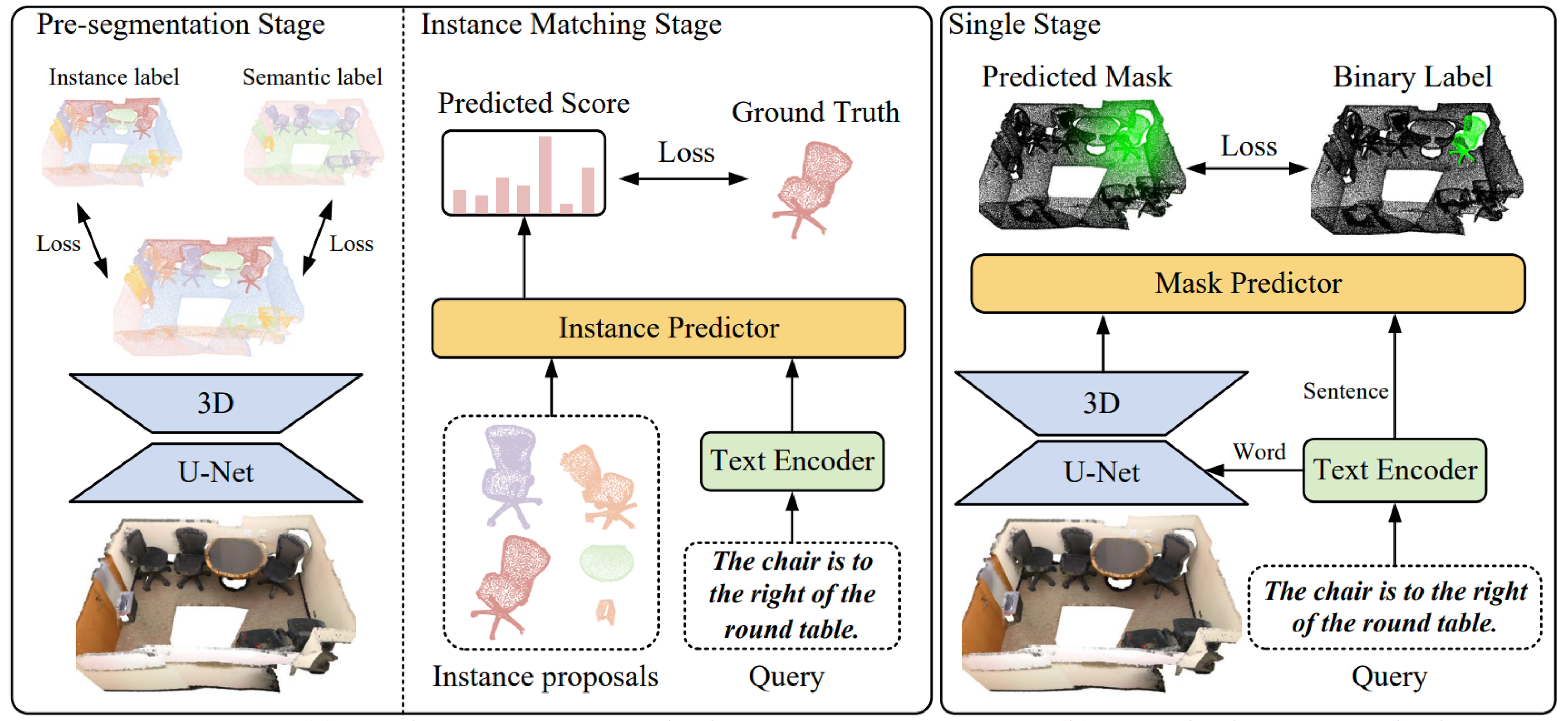
Xuexun Liu[1*] Xiaoxu Xu[1*] Jinlong Li[2*] Qiudan Zhang[1] Xu Wang[1†] Nicu Sebe[2] Lin Ma[3]

[1]Shenzhen University, [2]University of Trento, [3]Meituan Inc.

NEURAL INFORMATION PROCESSING SYSTEMS

Problem

- Previous referring 3D segmentation methods typically adopt segmentation-then-matching paradigm or utilize a powerful instance segmentation pre-train model as their backbone. These approaches all require both semantic and instance supervision signal.
- For previous segmentation-then-matching methods, target objects may be left out in the pre-segmentation stage because the network fails to focus on the objects that are more essential to the referring task.
- 3D scene is large and complex while the referred object is small. It is difficult to directly localize and segment target objects only with binary mask.

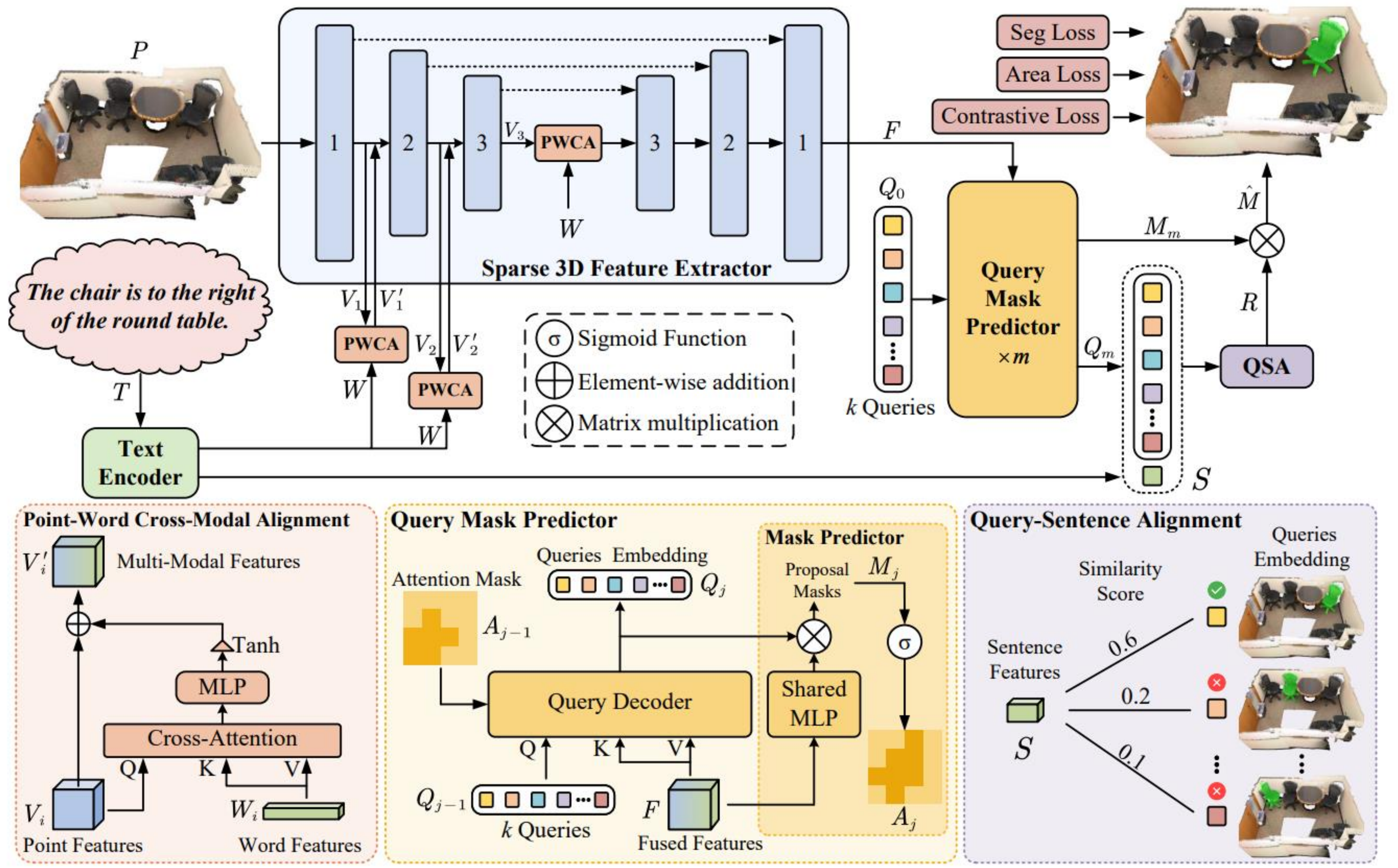


(a) Previous two-stage method (b) Our single-stage method

Contribution

- We propose a new Referring 3D Segmentation method, which directly performs referring 3D segmentation at a single stage to bridge the gap between detection and matching under the supervision of binary mask.
- To enhance cross-modal ability, we utilize a Point-Word Cross-Modal Alignment module and Query-Sentence Alignment module from coarse to fined.
- To reduce interference caused by multiple objects and backgrounds, we propose an area regularization loss and the point-to-point contrastive loss from coarse to fined.

Method



- **Area Regularization Loss:** Minimize the output probability of each point and promotes the network to predict a smallest mask.
- **Point-to-Point Contrastive Loss:** Pull the points from the described object together and push away the rest points.

$$\mathcal{L}_{area} = \frac{1}{N}\sum_{i=1}^{N}\sigma(\hat{M}_i) \qquad \mathcal{L}_{p2p} = -\frac{1}{|\mathcal{P}|}\sum_{i=1}^{|\mathcal{P}|}\frac{\exp(P_i\cdot P_{avg}/\tau)}{\exp(P_i\cdot P_{avg}/\tau)+\sum_{j=1}^{|\mathcal{N}|}\exp(P_i\cdot N_j/\tau)}$$

Time Consumption

Method	Inference (Whole Dataset) (min)	Inference (Per Scan) (ms)	Training (Stage 1) (h)	Training (Stage 2) (h)	Training (All) (h)
TGNN	27.98	176.57	156.02	8.53	164.55
X-RefSeg	20.00	126.23	156.02	7.59	163.61
Ours	**7.09**	**44.76**	-	-	**40.89**

Benchmark Results

- Comparison on Scanrefer dataset.

	Method	Backbone	Label Effort‡	Supervision	mIoU	Acc@0.25	Acc@0.5
Two Stage	TGNN	GRU	> 20 min	Ins.+ Sem.	26.10	35.00	29.00
	TGNN	BERT		Ins.+ Sem.	27.80	37.50	31.40
	X-RefSeg	GRU		Ins.+ Sem.	29.77	39.85	33.52
	X-RefSeg	BERT		Ins.+ Sem.	29.94	40.33	33.77
Single Stage	LESS (ours)	GRU	< 2 min	Mask	32.19	51.00	26.41
	LESS (ours)	BERT		Mask	32.44	51.41	29.02
	LESS (ours)	RoBERTa		Mask	**33.74**	**53.23**	**29.88**

‡ The evaluate of label effort is base on a single sample.

Ablation and Visualization

	PWCA	QSA	mIoU	A@25	A@50
(a)			32.66	51.71	27.20
(b)	✓		33.44	52.73	28.92
(c)	✓	✓	**33.74**	**53.23**	**29.88**

	$\mathcal{L}_{area}$	$\mathcal{L}_{p2p}$	mIoU	A@25	A@50
(a)			25.86	40.85	16.81
(b)	✓		31.04	49.61	24.72
(c)	✓	✓	**33.74**	**53.23**	**29.88**