Learning to Segment Every Referring Object Point by Point

Mengxue Qu[1,2*] Yu Wu [3] Yunchao Wei[1,2] Wu Liu [4] Xiaodan Liang [5,6] Yao Zhao[1,2†]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
[3]Wuhan University [4]JD Explore Academy [5]Sun Yat-sen University [6]MBZUAI

qumengxue@bjtu.edu.cn, wuyucs@whu.edu.cn, yunchao.wei@bjtu.edu.cn

1. Details of the Naive Pipeline

In the main paper, the naive pipeline for our proposed partially supervised RES is to first train a Referring Expression Comprehension (REC) model and then transfer it to the Referring Expression Segmentation (RES) task by fine-tuning on a limited number of data with mask annotations. In Fig. 1 of the main body, we compare the performance of the naive pipeline for the partially supervised setting in SeqTR [3] and MDETR [1] (dark green and dark blue curves in Fig. 1). We provide additional specifics here to demonstrate that our comparisons are reasonable, *i.e.*, MDETR lags behind SeqTR in partially supervised RES even though its fully supervised performance is better.

Specifically, we first train MDETR with SiRi [2] on full box-labeled data to obtain a REC model, which has a similar performance to that in SeqTR, and then transfer it to RES task by fine-tuning on RES data with mask annotations. For MDETR, a specific segmentation head is needed for RES, *e.g.*, a three-layer MLP. In addition, the loss function requires adjustment, *e.g.*, adding the pixel-level cross-entropy loss. For SeqTR, both the model and the loss function remain unchanged, the only thing that changes is the number of predicted coordinates. The predictions are four points (bounding box corners) in the REC task while dozens of points (contour) in the RES task. As shown in Fig. 1, when fine-tuning with full mask-labeled data, MDETR archives higher performance than SeqTR (light green and light blue curves). While with 1% mask-labeled data, SeqTR performs significantly better than MDETR, which means the sequence prediction model, *i.e.*, SeqTR, might be a better solution for partially supervised training.

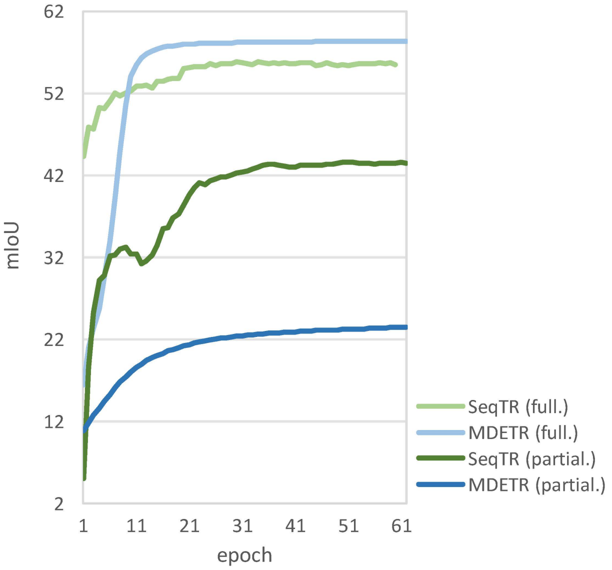

Figure 1. The IoU performance of the fully supervised SeqTR/MDETR and the naive partially supervised SeqTR/MDETR. Best viewed in color.

Method	[Task]	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
Multi-task	Task-wise	53.87	65.85	59.54	47.56	27.39	6.29
Refine REC	Zero	54.83	66.60	60.57	48.01	27.58	5.91
Refine REC	Learnable	**55.95**	**67.73**	**62.20**	**49.02**	**29.55**	**7.07**

Table 1. The impact of different [Task] token on *RefCOCOg@val* in fully-supervised RES.

2. Additional Experimental Analysis

Impact of the [Task] token. There is a [TASK] token before the sequence of points $\{x_i, y_i\}$. In SeqTR [3], it's set as a learnable vector for multi-task training or a vector with all zero values for single task REC/RES. In Table 1, we investigate the impact of different [Task] token on *RefCOCOg@val* in fully-supervised RES. In our method, we change the zero vector to learnable to better translate the coordinate representation in REC to RES.

[*] Work done during an internship at JD Explore Academy.
[†] Yao Zhao is the corresponding author.

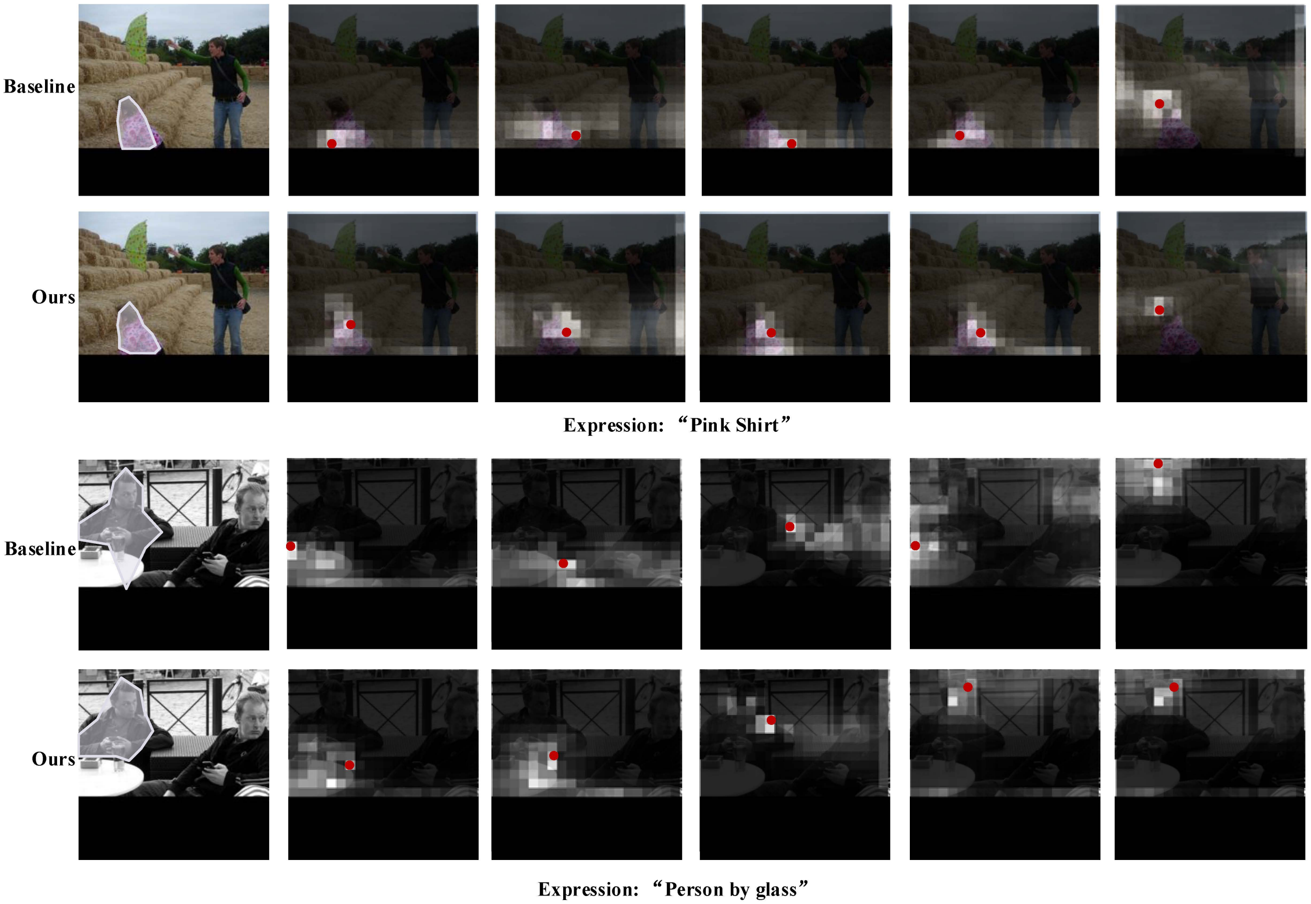

Figure 2. Visualization of Cross-Attention. The first column shows the prediction results of the baseline (the first row) and our method (the second row) when training with 1% mask-labeled data. For each sample, we randomly selected some contour coordinates (the points marked in red) to visualize their cross-attention map.

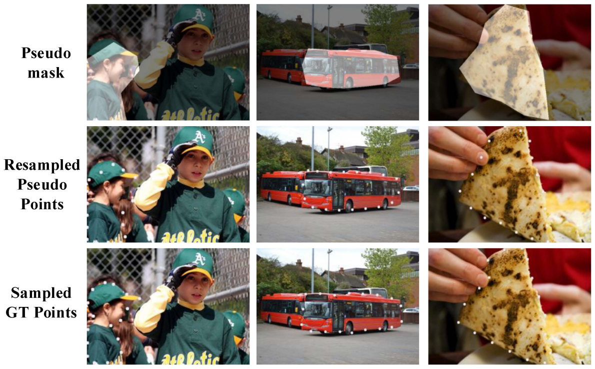

Figure 3. Visualization of the pseudo mask, uniformly re-sampled pseudo points, and the sampled ground truth (GT) points.

Visualization of the Re-sampled Pseudo Points We visualize the re-sampled pseudo points as shown in Fig. 3. The first row shows the binary mask generated by connecting the points predicted by the model, and the second row is the uniformly resampled pseudo points. Although the resampled point is still not accurate enough, it is consistent with the sampled ground truth point, which is uniform.

Visualization of Cross-Attention We further visualize the cross-attention map to reveal why our method can yield better contour points prediction in partially supervised RES in Fig. 2. The first column shows the contour prediction results of the baseline (the first row) and our method (the second row). For each sample, we randomly selected several coordinates (marked in red) to visualize their cross-attention map. As shown in Fig. 2, the brighter the area in the attention map denotes that the more attention it receives. With our proposed co-content teacher forcing (CCTF) and point-modulated cross-attention (PMCA), the perceptual focus of cross-attention is more focused on semantically relevant regions and filtering out some of the background noise.

References

[1] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-Modulated Detection for End-to-End Multi-Modal Understanding. In *ICCV*, 2021.

[2] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. Siri: A simple selective retraining mechanism for transformer-based visual grounding. In *ECCV*, pages 546–562. Springer, 2022.

[3] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, pages 598–615. Springer, 2022.