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

Abstract

Referring Expression Segmentation (RES) can facilitate pixel-level semantic alignment between vision and language. Most of the existing RES approaches require massive pixel-level annotations, which are expensive and exhaustive. In this paper, we propose a new partially supervised training paradigm for RES, i.e.*, training using abundant referring bounding boxes and only a few* (e.g., *1%) pixel-level referring masks. To maximize the transferability from the REC model, we construct our model based on the point-based sequence prediction model. We propose the co-content teacher-forcing to make the model explicitly associate the point coordinates (scale values) with the referred spatial features, which alleviates the exposure bias caused by the limited segmentation masks. To make the most of referring bounding box annotations, we further propose the resampling pseudo points strategy to select more accurate pseudo-points as supervision. Extensive experiments show that our model achieves 52.06% in terms of accuracy (versus 58.93% in fully supervised setting) on RefCOCO+@testA, when only using 1% of the mask annotations. Code is available at* https://github.com/qumengxue/Partial-RES.git.

1. Introduction

Referring Expression Segmentation (RES) aims to generate a segmentation mask for the object referred to by the language expression in the image. It allows pixel-level semantic alignment between language and vision, which is meaningful to many multi-modal tasks and can be applied to various practical applications, *e.g.*, video/image editing with sentences. Benefiting from the development of deep learning techniques, significant progress [31, 9, 12, 2, 19, 20, 27, 54, 9] has been made in this field and achieved remarkable performance.

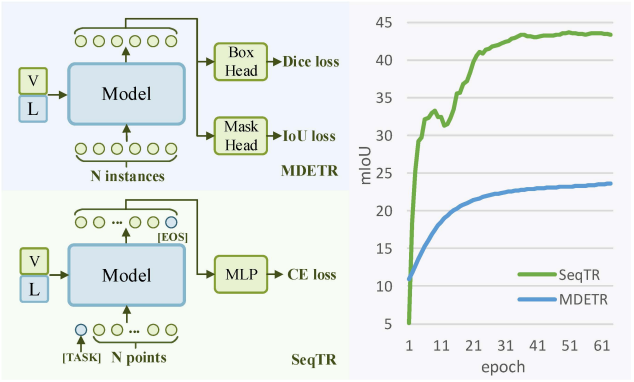


Figure 1. Comparison of MDETR and SeqTR. On the left is the streamlined framework of the two, and on the right is their IoU performance on *RefCOCOg@val*. Best viewed in color.

The success of current methods partially attributes to the large-scale training dataset with accurate pixel-level masks. However, labeling masks for a huge amount of images is often costly in terms of both human effort and finance, and thus hardly be scaled up. Therefore, it is meaningful to explore a new training paradigm for RES where extensive pixel-level annotations are not necessary. It is well known that bounding box annotations are much cheaper and easier to be collected compared with pixel-level masks. Thus, we try to study this question: *is it possible to learn an accurate RES model using abundant box annotations and only a few mask annotations?*

To answer this question, for the first time, we explore a new partially supervised training paradigm for RES (Partial-RES). Intuitively, a naive pipeline for this partially supervised setting is to first train a Referring Expression Comprehension (REC) model and then transfer it to the RES task by fine-tuning on a limited number of data with mask annotations. However, due to the difference between

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

the REC and RES tasks, prevalent models (*e.g.*, Mask R-CNN [14], MDETR [24]) have to simultaneously optimize two different heads, with one for detection prediction and the other for segmentation prediction. Such kind of structures would lead to a severe issue during the fine-tuning stage, *i.e.*, a randomly initialized mask head can not get well optimized given only a few mask annotations (hundreds of images). As shown in Figure 1, it is not easy to optimize MDETR with only 1% mask-annotated data even transferred from a well-trained REC model.

Recently, a contour-based method named SeqTR has been proposed for unifying REC and RES. The idea is to use a sequence model (usually a Transformer decoder) to sequentially generate contour points of the referred object. The predictions are two points (top-left and bottom-right corner points of the bounding box) in the REC task while dozens of points (contour) are in the RES task. As shown in Fig. 1, both boxes and masks are converted to point sequences and optimized using the same simple cross-entropy loss in SeqTR, which ensures the consistency between REC and RES. It incorporates two optimization heads into a unified one, making the knowledge learned with four points (*i.e.*, detection) be naturally transferred to predict multiple ones (*i.e.*, segmentation). Inspired by SeqTR, we have the conjecture that sequence prediction might be a better solution for partially supervised training.

In this paper, we investigate methods for achieving better partially supervised training for RES sequence prediction models. Sequence-to-sequence prediction models are typically trained with Teacher-Forcing (TF) [39], wherein the model utilizes the ground truth token as input to predict the next token during the training stage. However, the model can only predict the next state by taking its previous output as input in inference, which is known as the exposure bias [34]. This issue is even more severe in the partially supervised setting due to the very limited ground truth data. Consequently, we introduce the Co-Content Teacher-Forcing (CCTF), which combines the ground truth point coordinates together with the spatial visual feature of the pointed row or column. In contrast to the previous sequence model SeqTR, our CCTF explicitly associates the point coordinates (scale values) with the referred spatial region, providing a more natural approach for visual grounding.

Furthermore, we estimate the referring region via our proposed Point-Modulated Cross-Attention, to ensure the decoder attends to those region content while generating the point contour sequences. To fully utilize the data without mask annotation in Partial-RES, we retrain the model with the generated pseudo labels, and we present a Resampling Pseudo Points (RPP) Strategy. Unlike most pseudo-label works that directly use the network's predictions as labels, we select the appropriate pseudo masks with Dice coefficient and then resample the predicted points in a uniform way to regularize the contour sequence labels.

With extensive experiments, our method displays significant improvement compared to SeqTR [54] baselines on all three benchmarks, *i.e.*, at an average of 3.5%, 2.4%, and 3.0% on 1%, 5% and 10% mask-labeled data. With 10% mask annotated data, our method achieves 97% of the fully supervised performance on *RefCOCO+@val*. We are also able to achieve 88% of the fully supervised performance only with 1% mask-labeled data on *RefCOCO+@testA*.

2. Related Work

2.1. REC and RES

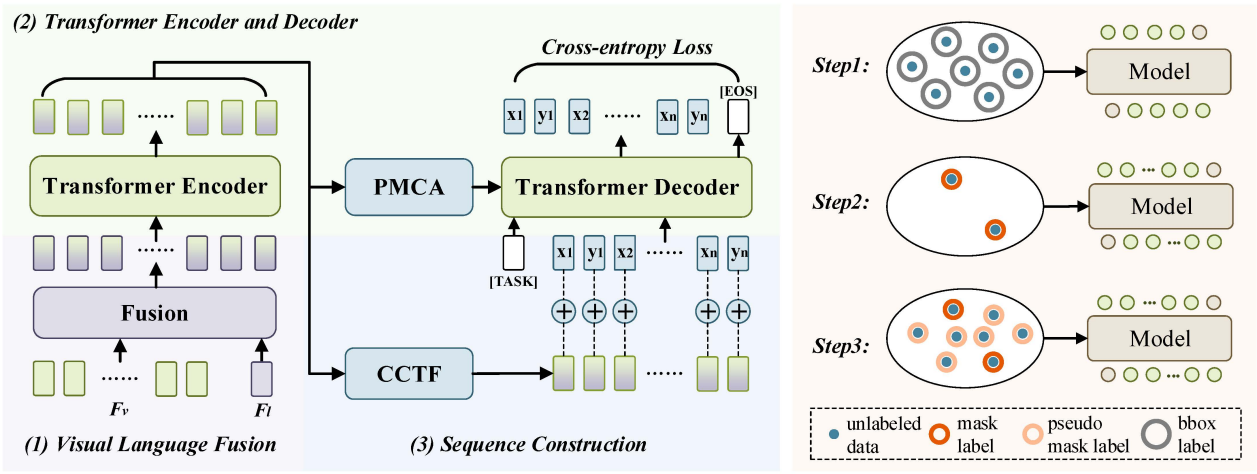
Referring Expression Comprehension (REC) aims at grounding the object referred to by a sentence, which achieves *instance-level* vision language alignment. Compared to REC, Referring Expression Segmentation (RES) grounds language expression at the fine-grained *pixel-level*. REC and RES are separate but closely related fundamental multi-modal tasks.

For REC, previous methods can be roughly divided into two-stage and one-stage. Two-stage methods [16, 18, 29, 40, 41, 43, 49, 52, 55] typically utilize an object detector to generate region proposals in the first stage, and are trained to maximize the similarity between region and text with binary cross-entropy loss or similarity loss. One-stage methods [5, 28, 38, 45, 46] avoid being constrained by the quality of the proposal by directly fusing visual and linguistic features rather than matching region-language pairs.

For RES, typical solutions are to incorporate as much linguistic information as possible in various ways on pixel-level visual features [31, 9, 12, 2, 19, 20]. In previous methods, researchers propose various attention mechanisms [9, 23, 20] to better merge vision and language. Recently, since the superior ability of modeling vision and language of Transformer-based model, more high-performance methods [27, 54, 9, 36] are proposed to solve RES tasks.

For REC and RES, multi-task methods [32, 27] are proposed to better use the correlation between two tasks, which can facilitate learning of both tasks. However, all the prior work require different task-specific branch and loss function, whose generalization ability is limited. Recently, a simple and universal network term SeqTR [54] is presented, which regards both REC and RES as a point prediction problem. SeqTR greatly reduces the difficulty and complexity of both architecture design and optimization.

In some work, REC is treated as a pretext task for other multi-modal tasks [24, 51], *e.g.*, Visual Question and Answering. RES has a more fine-grained visual language alignment capability than REC, but the capability of RES to be used in multi-modal pre-training has not yet been developed because of the lack of a large-scale RES dataset. It is well known that annotating binary masks for a multi-



(a) Architecture (b) Training Steps

Figure 2. (a) Overall architecture of our method. It consists of three main modules: (1) Visual-Language Fusion; (2) Transformer Encoder and Decoder; (3) Sequence Construction. "CCTF": Co-Content Teacher-Forcing; "PMCA": Point-Modulated Cross-Attention. (b) The suggested training steps of our method in the Partially Supervised RES task. Best viewed in color.

modal dataset is expensive while bounding box is easy. In this paper, we make use of the bounding box annotations and the consistent task expression of REC and RES in SeqTR, attempting to train RES task with less mask-labeled data based on SeqTR.

2.2. Partially Supervised Instance Segmentation

Partially Supervised Instance Segmentation (PSIS) was first raised in [17], where object categories are divided into *base* and *novel* splits. Both of them have bounding box annotation while only *base* have mask annotation. The training target of this task is to generate segmentation for *novel* categories. PSIS has been studied for a period of time in the field of Instance Segmentation [1, 11, 26, 35, 53]. Some works address this problem by learning class-agnostic mask segmentation models, *i.e.*, separate foreground and background only. They capture class-agnostic cues, *e.g.*, shape [26] and appearance commonalities [11]. For better use of the training data from *novel* categories, [42] propose ContrastMask to learn a mask segmentation model on both base and novel categories under a unified pixel-level contrastive learning framework. In this paper, we suggest the partially supervised RES, drawing on the partially supervised instance segmentation. We train the RES model with bounding box annotation of all *train* data, and a small fraction of mask annotation.

2.3. Sequence Prediction in Visual Tasks

Visual tasks differ greatly in their output format and associated content, making it difficult to handle them with an identical structure. Unlike previous approaches requiring prior knowledge in Object Detection, Pix2Seq [4] first cast object detection as a language modeling task conditioned on the observed pixel inputs. The class and the bounding box are expressed as sequences of discrete tokens, and train the detection model to generate the coordinate and class of each tone by one. Built on Pix2Seq, [6] proposes an object-centric vision framework Obj2Seq that takes objects as the basic unit, and human pose estimation is also been converted to a sequence-generated form. In addition, another similar work SeqTR [54] is proposed to unify REC and RES, regarding both REC and RES as a point prediction problem. Inspired by the above-mentioned work, we consider whether this sequence prediction framework can be used in scenarios with few labels to take advantage of its good format uniformly property.

3. Method

In this section, we first briefly revisit the sequence prediction model SeqTR in Sec. 3.1. Then we illustrate the architecture of our method in Sec. 3.2. In Sec. 3.3, we elaborate on our proposed Co-Content Teacher-Forcing, and the Point-Modulated Cross Attention in Sec. 3.4 and the Resampling Pseudo Point Strategy in Sec. 3.5.

3.1. Revisit SeqTR

SeqTR [54] is a sequence prediction model for Visual Grounding tasks, which first reformulate visual grounding as a point prediction problem. The bounding box and the segmentation mask are serialized into a sequence of discrete coordinate tokens $\{x_i, y_i\}_{i=1}^{N}$, where N is 4 vertices of the bounding box on REC or the number of contour

points of the segmentation mask on RES. With a task token [TASK], SeqTR can predict the target coordinate tokens in an auto-regressive manner during inference, and end the coordinate sequence with [EOS] token. In addition, different from those loss functions adopted in DETR-like multi-task frameworks (*e.g.*, GIoU loss, set-based matching loss, focal loss, and dice loss), SeqTR only uses a simple cross-entropy loss for REC and RES, which requires no further prior knowledge or expertise. SeqTR greatly reduces the difficulty and complexity of both architecture design and optimization for visual grounding tasks. Meanwhile, it nicely aligns REC and RES, and naturally converts them into one type of expression form. Following SeqTR, our model is also optimized with a simple cross-entropy loss.

3.2. Architecture

We build our architecture on the latest RES sequence prediction method SeqTR [54], which consists of three main modules: (1) Visual-Language Fusion; (2) Transformer Encoder and Decoder; (3) Sequence Construction; The overall architecture is illustrated in Fig. 2.

Visual-Language Fusion. We adopt the convolutional backbone DarkNet-53 [37] to obtain the visual representation for an input image $\mathbf{I}$. Following SeqTR, we downsample the multi-scale visual feature from the finest to coarsest spatial resolution and flatten it to $F_v \in R^{(H\times W)\times C}$. For referring expressions, we encode the text with a one-layer bidirectional GRU [7] as in SeqTR. Then we concatenate both unidirectional hidden states $h_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right]$ at each step t to get language feature $F_l = \{h_t\}_{t=1}^{T}$. Before V-L fusion, we adopt max pooling $mp(\cdot)$ along the channel dimension of F_i. Then visual feature F_v and language feature F_l will be fused to F_m by Hadamard product:

$$F_m = \delta(F_v) \odot \delta(mp(F_l)), \tag{1}$$

here the activation function $\delta(\cdot)$ is set to $tanh(\cdot)$ if no special instructions.

Transformer Encoder and Decoder. We use the Transformer with 6 layers encoder and 3 layers decoder to learn the multi-modal feature and decode out the vertices or the contour points. The hidden dimension of transformer is set to 256. The fused multi-modal feature F_m is added sine and learned positional encoding [39] before entering Transformer Encoder. The original Cross-Attention between Transformer Encoder and Decoder will be replaced by Point-Modulated Cross-Attention (PMCA), which is introduced in Sec. 3.4.

Sequence Construction. Following SeqTR [54], we construct the sequence of floating points $\{\tilde{x_i}, \tilde{y_i}\}_{i=1}^{N}$ with the top-left and bottom-right corner points of the bounding box (N=2). For the binary mask on RES, the sequence is uniformly sampled N points clockwise on top of the mask on RES. Then they are quantized into integer bins by

$$x_i = round(\frac{M\tilde{x_i}}{w}),\ y_i = round(\frac{M\tilde{y_i}}{h}) \tag{2}$$

where each coordinate is normalized by image width w and height h, and M is the number of quantization bins.

The sequence will be embedded and then combined with the output multi-modal feature of the Transformer Encoder to conduct Co-Content Teacher-Forcing (CCTF) training. Details are described in Sec. 3.3. Moreover, we will resample pseudo points and reconstruct the sequence for model retraining. Details can be found in Sec. 3.5.

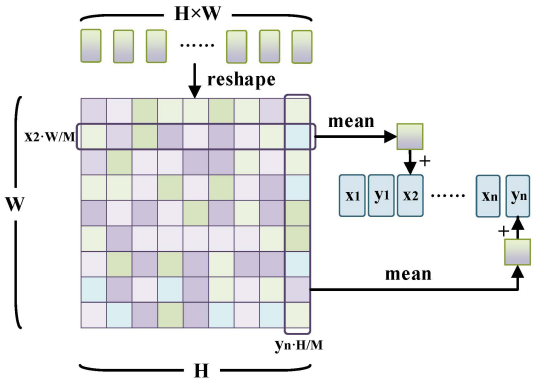


Figure 3. Sketch of CCTF. We extract multi-modal features from the reshaped Transformer Encoder output according to the target coordinates and fuse them by element-wise addition, *e.g.*, x_n corresponds to the x_nW/M line of row of the feature map, and y_n corresponds to the y_nH/M line of column.

3.3. Co-Content Teacher-Forcing

The Co-Content Teacher-Forcing is proposed to alleviate the over-reliance on the sampled ground truth points during training, which leads to exposure bias of the inconsistency of Training and Inference.

In the normal Teacher-Forcing training, the sequence $\{x_i, y_i\}_{i=1}^{N}$ will be embed to a feature sequence $\{m_i, n_i\}_{i=1}^{N}$:

$$\{m_i, n_i\}_{i=1}^{N} = dic(\{x_i, y_i\}_{i=1}^{N}), \tag{3}$$

where $dic(\cdot)$ is a learnable coordinate embedding dictionary. The coordinate sequence $\{x_i, y_i\}_{i=1}^{N}$ can be viewed as the index of the dictionary. However, in the Co-Content Teacher-Forcing, the feature sequence will be combined with the encoder output multi-modal feature. As shown in Fig. 3, we reshape the output multi-modal feature $F_m \in R^{H\times W\times C'}$ of Transformer Encoder to $F_m{}' \in R^{(H\times W)\times C'}$ for better correspondence with coordinate points. Then, for each coordinate x_i or y_i, we extract the feature of the corresponding row or column from the feature map $F_m{}'$ and average them with $E\{\cdot\}$:

$$f_{x_i} = E_{k=1}^{H}\{F_m{}'(\frac{W}{M}x_i, k)\},$$
$$f_{y_i} = E_{k=1}^{W}\{F_m{}'(k, \frac{H}{M}y_i)\}, \tag{4}$$

M is the number of quantization bins. $\{\frac{W}{M}x_i, \frac{H}{M}y_i\}_{i=1}^{N}$ is the normalized index. Then the sequential feature $\{m_i, n_i\}_{i=1}^{N}$ will be combined with $\{f_{x_i}, f_{y_i}\}_{i=1}^{N}$ by simple element-wise addition, the final sequence input to Decoder is represented as

$$S = \{m_i, n_i\}_{i=1}^{N} + \{f_{x_i}, f_{y_i}\}_{i=1}^{N} \tag{5}$$

With CCTF training, our model explicitly associate the point coordinates (scale values) with the referred spatial region, which is more natural for visual grounding.

3.4. Point-Modulated Cross Attention

We adopt the SMCA proposed in [13] to replace the original Cross-Attention, which can constrain the cross-attention responses via estimating the Gaussian-like referring region prior. In the sequence prediction model, we generate the attention map for each co-content coordinates embedding, leading to a matched cross-attention region learned when the supervised data is reduced. To distinguish we denote it as Point-Modulated Cross-Attention (PMCA).

For each co-content coordinate embedding $S = \{f_j\}_{j=1}^{2N} = \{f_{x_i}, f_{y_i}\}_{i=1}^{N}$, we use a 3-layer MLP followed by a sigmoid activation function to generate a prior Gaussian-like attention map. The center point coordinates c_w, c_h and the scales s_w, s_h of the prior Gaussian map is denoted as

$$s_w, s_h, c_w, c_h = sigmoid(MLP(f_j)), \tag{6}$$

and the Gaussian map is represented as

$$G(u, v) = exp\left(-\frac{(u - c_w)^2}{\sigma s_w^2} - \frac{(v - c_h)^2}{\sigma s_h^2}\right) \tag{7}$$

where $u \in R^W, v \in R^H$, they are the spatial indices of G, and σ is a hyper-parameter to modulate the bandwidth of the Gaussian map.

3.5. Resampling Pseudo Point Strategy

To fully utilize the data without mask annotation in Partial-RES, we retrain the model with the generated pseudo label, and we propose a Resampling pseudo point strategy. To remain the consistency of the pseudo points with the original sampled contour points of the ground truth mask, we will resample the pseudo points. As shown in Fig. 4, we first connect the points predicted by the network to a binary mask. Then we compute the dice coefficient D of the projection of this mask and the ground truth bounding box.

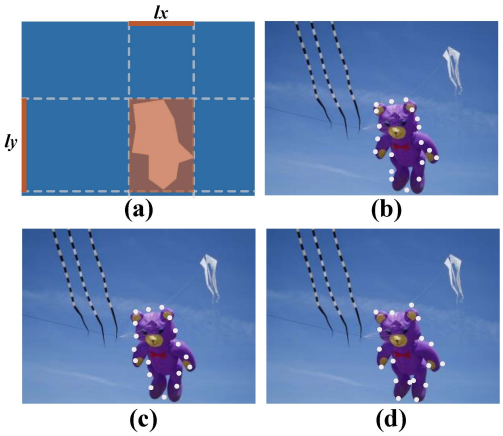


Figure 4. Sketch of Resampling Pseudo Point (RPP) Strategy. (a) Screening the appropriate pseudo mask with the dice coefficient between the predicted binary mask and the ground truth bounding box; (b) The generated points by the RES model; (c) The uniformly resampled pseudo points from pseudo mask; (d) The uniformly sampled target points from ground truth mask.

$$D(l, b) = dice(l_y, b_y)) + dice(l_x, b_x)) \tag{8}$$

where l_x, l_y are the projection of the mask on x-axis, y-axis, b_x, b_y are of the ground truth bounding box.

We set a threshold T for $D(l, b)$ to pick out those pseudo masks that are more accurate to avoid introducing too much noise by pseudo labels. For the selected samples, the predicted points are spaced may be inhomogeneous as shown in Fig. 4(b). So we resample the pseudo points with a uniform sampling scheme. It is considered that the contour points of previously labeled data are sampled uniformly and the pseudo point should be kept consistent with it.

We propose a feasible training process for the Partially Supervised RES task as shown in Fig. 2 (b):

Step1. REC Training. Since in the Partial-RES task, all the bounding box labels are seen in the model training, we first train the model on fully supervised REC to get a better representation relatively of multi-model features from the encoder output. The feature will be used to enhance TF learning in the next step of RES training. In practice, we load the REC model trained in SeqTR as a pretrained one.

Step2. Initial RES Training. In this step, we train the model with only 1%, 5%, and 10% labeled mask data on the RES task. The model is equipped with the new method proposed in this paper, including CCTF and PMCA.

Step3. Retraining with Pseudo Points. After obtaining the RES model from *Step2*, we can utilize it to generate more pseudo contour points. While these points may contain a lot of noise or may not match the way labeled data is sampled, so we re-selected and sampled out the appropriate pseudo points with the proposed RPP strategy.

Label Ratio	Method	Venue	Visual Encoder	RefCOCO			RefCOCO+			RefCOCOg		
				val	testA	testB	val	testA	testB	val-g	val-u	test-u
100%	*Fully-supervised:*											
	CMSA [47]	CVPR'19	RN101	58.32	60.61	55.09	43.76	47.60	37.89	39.98	-	-
	STEP [3]	ICCV'19	RN101	60.04	63.46	57.97	48.19	52.33	40.41	46.40	-	-
	CMPC [21]	CVPR'20	DN53	61.36	64.53	59.64	49.56	53.44	43.23	49.05	-	-
	LSCM [22]	ECCV'20	DLA34	61.47	64.99	59.55	49.34	53.12	43.50	48.05	-	-
	MCN [32]	CVPR'20	DN53	62.44	64.20	59.71	50.62	54.99	44.69	-	49.22	49.40
	BUSNet [44]	CVPR'21	RN101	63.27	66.41	61.39	51.76	56.87	44.13	50.56	-	-
	LTS[21] [23]	CVPR'21	DN53	65.43	67.76	63.08	54.21	58.32	48.02	-	54.40	54.25
	VLT [9]	ICCV'21	DN56	65.65	68.29	62.73	55.50	59.20	49.36	49.76	52.99	56.65
	ResTR [25]	CVPR'22	ViT-B	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-
	SeqTR [54]	ECCV'22	DN53	67.26	69.79	64.12	54.14	58.93	48.19	-	55.67	55.64
10%	*Partially-supervised*											
	Baseline	-	DN53	57.92	60.74	55.06	50.84	54.87	44.54	-	49.50	49.30
	Our Method	-	DN53	**63.99**	**65.86**	**61.54**	**52.56**	**55.79**	**46.09**	-	**52.43**	**52.40**
5%	Baseline	-	DN53	58.85	61.72	56.36	49.78	54.22	43.34	-	48.33	48.27
	Our Method	-	DN53	**62.23**	**64.67**	**60.42**	**51.30**	**54.68**	**44.85**	-	**51.28**	**50.55**
1%	Baseline	-	DN53	54.67	57.48	52.62	44.91	48.81	39.44	-	42.57	42.46
	Our Method	-	DN53	**57.41**	**59.60**	**55.84**	**47.68**	**52.06**	**41.54**	-	**46.80**	**46.12**

Table 1. Comparisons with state-of-the-art fully-supervised methods on RefCOCO [50], RefCOCO+ [50], and RefCOCOg [33] in terms of IoU scores. We defined mask label ratio as 1%, 5%, and 10% to compare our method with the baseline in partially-supervised RES. "RN101": ResNet-101 [15], "DN53": DarkNet-53 [37], "DN56": DarkNet-56 [37], "DLA-34" [48], "ViT-B" [10].

4. Experiments

4.1. Datasets

RefCOCO/RefCOCO+ are proposed in [50]. There are 19,994 images in RefCOCO with 142,209 refer expressions for 50,000 objects. Similarly, 19,992 images are included in RefCOCO+ which contains 141,564 expressions for 49,856 objects.

In these datasets, each image contains two or more objects from the same category. In RefCOCO+ dataset, positional words are not allowed in the referring expression, which is a pure dataset with appearance-based referring expression, whereas RefCOCO imposes no restriction on the phrase. In addition to the training set and validation set, the test set for RefCOCO/RefCOCO+ is divided into a testA set (containing several people in an image) and a testB set (containing multiple instances of other objects in an image).

RefCOCOg [33] contains 26,711 images with 85,474 referring expressions for 54,822 objects, and each image usually contains $2 \sim 4$ objects of the same category. The length of referring expressions in this dataset is almost twice as long as those in RefCOCO and RefCOCO+. This dataset is directly divided into "train", "val" and "test".

4.2. Experimental Settings

Each instance to be segmented is equipped with an expression sentence in RES datasets, the sentence can be seen as a kind of open category. Therefore, different from the manner of dividing data [17], we simulate the partially supervised training scenario on RES by randomly sampling the entire train data proportionally without considering the category.

Implementation Details. Following SeqTR [54], all parameters in the network are optimized with AdamW [30], and the batch size is 128. We train the model with $120 \sim 150$ epochs for full convergence, and we set the learning rate as 5e-4@10%, 3e-4@5%, and 1e-4@1% in *step2* training, while 5e-4 in *step3* training with pseudo points. We use DarkNet-53 [37] as the visual encoder. Following standard practices [8, 32, 23, 9], images are resized to 640×640, and the length of language expression is trimmed at 15 for ResCOCO/RefCOCO+ and 20 for RefCOCOg. The training epochs and the learning rate used vary slightly for different settings of mask-labeled data ratio and different training stages. For the number of sampled contour points, we use the same as in SeqTR, *i.e.*, 18 points for RefCOCO/RefCOCO+, and 12 for RefCOCOg.

Evaluation Metrics. Following the proposal setting in the previous work, we use mask Intersection-over-Union (IoU) and Precision with thresholds (Pr@X). The mask IoU demonstrates the mask quality, which emphasizes the model's overall performance and reveals both targeting and segmenting abilities. The Pr@X metric computes the ratio of successfully predicted samples using different IoU thresholds. Low threshold precision like Pr@0.5 reflects the identification performance of the method, and high threshold precision like Pr@0.9 reveals the ability to generate high-quality masks.

Ratio	CCTF	PMCA	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
1%			42.57	47.54	32.89	17.61	5.07	0.36
	✓		44.20 (+1.63)	50.68 (+3.14)	37.00 (+4.11)	21.21 (+3.60)	7.89 (+2.82)	0.66 (+0.30)
	✓	✓	**44.58 (+0.38)**	**50.98 (+0.30)**	**37.32 (+0.32)**	**22.42 (+1.21)**	**8.21 (+0.32)**	**0.76 (+0.10)**
100%			55.95	67.73	62.20	49.02	29.55	7.07
	✓		56.03 (+0.08)	67.79 (+0.06)	62.32 (+0.12)	49.88 (+0.86)	30.63 (+1.08)	8.04 (+0.97)
	✓	✓	**56.89 (+0.86)**	**68.43 (+0.64)**	**62.96 (+0.64)**	**50.72 (+0.84)**	**31.21 (+0.58)**	**8.05 (+0.01)**

Table 2. Validity verification of our proposed CCTF and PMCA. We conduct experiments on 1% and 100% supervised data with mask labels (100% means fully-supervised training).

Method	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
Baseline	44.58	50.98	37.32	22.42	8.21	0.76
w/o RPP	44.27	51.28	38.24	24.28	10.24	0.96
w/ RPP	**46.80**	**56.29**	**43.19**	**27.40**	**10.34**	**0.98**

Table 3. Comparison of our method with re-sampling pseudo points and without RPP in *step3* training.

Method	IoU	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
Baseline	44.58	50.98	37.32	22.42	8.21	0.76
T@1.00	44.53	51.36	38.14	22.40	7.29	0.58
T@0.05	44.35	50.68	38.52	24.18	9.70	0.78
T@0.10	**46.80**	**56.29**	**43.19**	**27.40**	**10.34**	**0.98**

Table 4. Performance with different Threshold T of dice coefficient on *RefCOCOg@val*. T@1.00 means that all the predicted pseudo masks are used for training in *step3* without selection of the more accurate one with dice coefficient.

Baseline and oracle We suggest the baseline in the following way: we initialize our model with the parameters of SeqTR trained on the REC task, and refine it on the RES task with 1%, 5%, and 10% mask annotation data. Our comparison with the baseline is discussed in Sec. 4.3. The "oracle" model is SeqTR trained on REC and refined on RES with all mask annotation data. This fully supervised model is a performance upper bound for our partially supervised RES.

4.3. Comparison with other Method

We compare our method with other fully supervised state-of-the-art methods on three common benchmarks of Referring Expression Segmentation, *i.e.*, RefCOCO, RefCOCO+, and RefCOCOg. In addition, we also compare with our suggested baseline in partially-supervised training. Results are reported in Table 1. Our method displays significant improvement over the baseline method on all three datasets, at an average of 3.5%, 2.4%, and 3.0% on 1%, 5%, and 10% mask labeled data. With 10% mask annotated data, we can achieve 97% of the fully supervised performance on *RefCOCO+@val*. We are also able to achieve 88% fully supervised performance on 1% mask-labeled data on *RefCOCO+@testA*. Our methods can even surpass some fully-supervised methods, *e.g.*, MCN, [32], BUSNet [44].

4.4. Ablation Studies

Effectiveness of CCTF and PMCA. To verify the effectiveness of Co-Content Teacher Forcing (CCTF) and Point-Modulated Cross-Attention (PMCA), we show the results of our RES model with 1% partially supervised training data and 100% fully supervised training data on RefCOCOg@val in Table 2.

When trained with 100% mask label of the data, PMCA embodies more performance enhancements than CCTF, which proves the effectiveness of PMCA. CCTF contributes little to performance improvement. While with 1% supervised mask labeled training data, the improvement contribution of CCTF far exceeds that of PMCA, 80% performance improvement comes from CCTF (1.63% in 2.01%). This demonstrates that the exposure bias of the inconsistency of the training and inference process may be more serious in the case of partially supervised data, which is alleviated by our proposed CCTF somewhat.

Ablation of Resampling Pseudo Points To better demonstrate the effectiveness of re-sampling pseudo points (RPP), we compare the results with those obtained without re-sampling. In the case of re-sampling, we uniformly re-sample the pseudo points from the generated pseudo mask and consider them as the target sequence. In contrast, for no re-sampling, we use the points predicted by the model directly as the target sequence. As illustrated in Table 3, the reported performance is superior when the RPP strategy is employed. This is primarily attributed to the fact that after re-sampling, the unlabeled data is pseudo-labeled uniformly, which mitigates the potential noise introduced by pseudo labels.

Performance with different Threshold T of dice coefficient. We establish a threshold for the dice coefficient to select more suitable pseudo points that will serve as prediction targets for unlabeled mask annotation data in *step3* mentioned in Sec. 3.5. To demonstrate the effectiveness of our pseudo points selection strategy and the impact of different threshold, we compare the results with different thresholds T of the dice coefficient on *RefCOCOg@val*. As presented

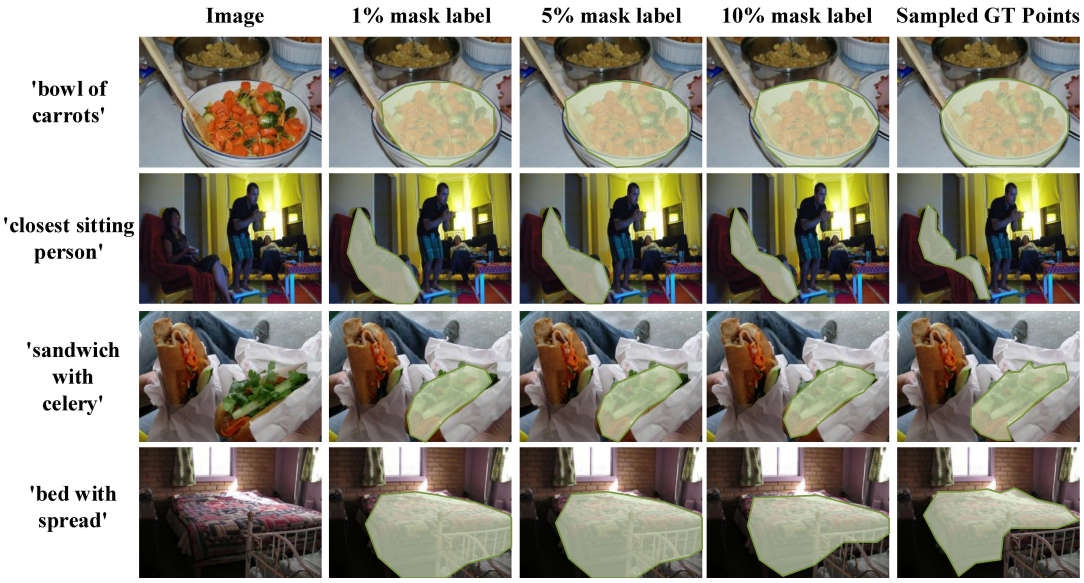


Figure 5. Qualitative examples of our method at 1%, 5%, 10% mask label in Partially Supervised RES. We identify the image and its corresponding referring expression in the first column.

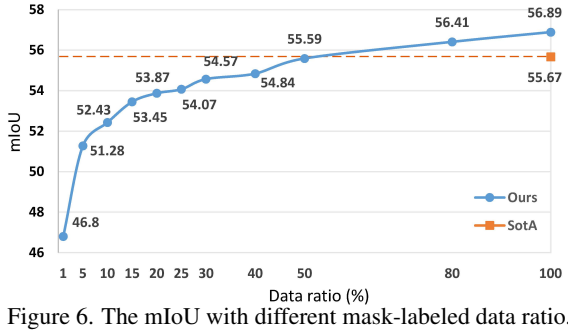


Figure 6. The mIoU with different mask-labeled data ratio.

in Table 4, without any constraint (T is set to 1.00), we observe that the side effect of pseudo points is greater than the main effect due to the noise contained in pseudo points. Therefore, we determine to select out those pseudo masks with *dice coefficient* $< T$ for re-sampling pseudo points. Meanwhile, our ablation studies declare that the threshold of 0.10 is the optimal choice for our selection strategy.

Experiment of increasing the proportion of mask-labeled data We conduct an additional experiment to investigate the proportion of mask-labeled data in training that can reach SotA performance. As depicted in Fig. 6, comparing the mIoU with the fully supervised SotA method (55.67) on *RefCOCOg@val*, our method attains remarkably similar results (55.59) by using 50% mask-labeled data.

4.5. Qualitative Results

We evaluated the segmentation performance of our method at various ratios, *i.e.*, 1%, 5%, and 10% of Partial-RES data in Fig. 5. As the mask-labeled data increases, the contours fit more accurately. By comparing these results, we observe that with a low percentage of mask-labeled data, our segmentation results are already similar to those of the GT points.

5. Conclusion

In this paper, we investigate the partially supervised learning paradigm for RES, where only a few segmentation masks are available so the model has to learn transfer knowledge from REC. Unlike existing RES/REC works that have different decoders for the two tasks, we leverage the contour-based sequence prediction model to maximize the transfer ability. We then propose CCTF to explicitly associate the point coordinates (scale values) with the referred spatial regions, which alleviates the exposure bias brought by the limited segmentation masks. Besides, we present RPP strategy to select appropriate pseudo labels and resample pseudo points during the retraining stage. We conduct extensive experiments to demonstrate the effectiveness of our model in the Partial-RES setting. One limitation of this work is that the smoothness and edge accuracy of the point-contour mask prediction may not be as precise as the pixel-level mask predictions.

Acknowledgements. This work was supported by the National Key RD Program of China (No.2018AAA0102100), the National Natural Science Foundation of China (Nos. U1936212, 62120106009), Sponsored by Beijing Nova Program (NO. 20220484063).

References

[1] David Biertimpel, Sindi Shkodrani, Anil S Baslamisli, and Nóra Baka. Prior to segment: Foreground cues for weakly annotated classes in partially supervised instance segmentation. In *ICCV*, pages 2824–2833, 2021.

[2] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, pages 7454–7463, 2019.

[3] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, pages 7454–7463, 2019.

[4] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*.

[5] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. Real-time Referring Expression Comprehension by Single-stage Grounding Network. *arXiv preprint arXiv:1812.03426*, 2018.

[6] Zhiyang Chen, Yousong Zhu, Zhaowen Li, Fan Yang, Wei Li, Haixin Wang, Chaoyang Zhao, Liwei Wu, Rui Zhao, Jinqiao Wang, et al. Obj2seq: Formatting objects as sequences with class prompt for visual tasks. In *NeurIPS*.

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[8] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding with Transformers. *ICCV*, 2021.

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[11] Qi Fan, Lei Ke, Wenjie Pei, Chi-Keung Tang, and Yu-Wing Tai. Commonality-parsing network across shape and appearance for partially supervised instance segmentation. In *ECCV*, pages 379–396. Springer, 2020.

[12] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, pages 15506–15515, 2021.

[13] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *ICCV*, pages 3621–3630, 2021.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[16] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[17] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, pages 4233–4241, 2018.

[18] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *CVPR*, 2017.

[19] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020.

[20] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020.

[21] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020.

[22] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75. Springer, 2020.

[23] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021.

[24] 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.

[25] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18145–18154, 2022.

[26] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *ICCV*, pages 9207–9216, 2019.

[27] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34:19652–19664, 2021.

[28] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A Real-time Cross-modality Correlation Filtering Method for Referring Expression Comprehension. In *CVPR*, 2020.

[29] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *ICCV*, 2019.

[30] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam. 2018.

[31] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *ACM MM*, pages 1274–1282, 2020.

[32] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *CVPR*, 2020.

[33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 2016.

[34] Tsvetomila Mihaylova and André FT Martins. Scheduled sampling for transformers. *ACL 2019*, page 351, 2019.

[35] Pedro O O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *NeurIPS*, 28, 2015.

[36] 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.

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[38] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-Shot Grounding of Objects from Natural Language Queries. In *ICCV*, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeuruIPS*, 2017.

[40] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[41] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Aetworks. In *CVPR*, 2019.

[42] Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, and Wei Shen. Contrastmask: Contrastive learning to segment every thing. In *CVPR*, pages 11604–11613, 2022.

[43] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic Graph Attention for Referring Expression Comprehension. In *ICCV*, 2019.

[44] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *CVPR*, pages 11266–11275, 2021.

[45] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving One-Stage Visual Grounding by Recursive Sub-Query Construction. In *ECCV*, 2020.

[46] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A Fast and Accurate One-Stage Approach to Visual Grounding. In *ICCV*, 2019.

[47] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019.

[48] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, pages 2403–2412, 2018.

[49] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*, 2018.

[50] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016.

[51] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, pages 11782–11791, 2021.

[52] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding Referring Expressions in Images by Variational Context. In *CVPR*, 2018.

[53] Yanzhao Zhou, Xin Wang, Jianbin Jiao, Trevor Darrell, and Fisher Yu. Learning saliency propagation for semi-supervised instance segmentation. In *CVPR*, pages 10307–10316, 2020.

[54] 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.

[55] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel Attention: A Unified Framework for Visual Object Discovery Through Dialogs and Queries. In *CVPR*, 2018.