Rethinking Two-Stage Referring Expression Comprehension: A Novel Grounding and Segmentation Method Modulated by Point

Peizhi Zhao[1], Shiyi Zheng[1], Wenye Zhao[1], Dongsheng Xu[1], Pijian Li[1], Yi Cai[3,4], Qingbao Huang[1,2]*

[1]School of Electrical Engineering, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
[3]School of Software Engineering, South China University of Technology, Guangzhou, China
[4]Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
2212391086@st.gxu.edu.cn, qbhuang@gxu.edu.cn

Abstract

As a fundamental and challenging task in the vision and language domain, Referring Expression Comprehension (REC) has shown impressive improvements recently. However, for a complex task that couples the comprehension of abstract concepts and the localization of concrete instances, one-stage approaches are bottlenecked by computing and data resources. To obtain a low-cost solution, the prevailing two-stage approaches decouple REC into localization (region proposal) and comprehension (region-expression matching) at region-level, but the solution based on isolated regions cannot sufficiently utilize the context and is usually limited by the quality of proposals. Therefore, it is necessary to rebuild an efficient two-stage solution system. In this paper, we propose a point-based two-stage framework for REC, in which the two stages are redefined as point-based cross-modal comprehension and point-based instance localization. Specifically, we reconstruct the raw bounding box and segmentation mask into center and mass scores as soft ground-truth for measuring point-level cross-modal correlations. With the soft ground-truth, REC can be approximated as a binary classification problem, which fundamentally avoids the impact of isolated regions on the optimization process. Remarkably, the consistent metrics between center and mass scores allow our system to directly optimize grounding and segmentation by utilizing the same architecture. Experiments on multiple benchmarks show the feasibility and potential of our point-based paradigm. Our code available at https://github.com/VILAN-Lab/PBREC-MT.

Introduction

Referring Expression Comprehension (REC) aims to predict a referred target in an image according to a corresponding expression, which can be regarded as a coupling of the comprehension of visual and linguistic abstract concepts and the localization of concrete visual instance. Depending on the prediction paradigm, the comprehension of the referring expression has two manifestations: 1) Referring Expression Grounding (REG) (Yu et al. 2016; Mao et al. 2016), where the referred instance is localized by a bounding box. 2) Referring Expression Segmentation (RES) (Hu, Rohrbach, and Darrell 2016; Liu et al. 2017) separates the foreground and background of the image based on the referring expression. As a fundamental cross-modal task, REC focuses on mining fine-grained visual and linguistic information, which facilitates numerous downstream studies, such as autonomous driving (Kim et al. 2019), image captioning (Chen et al. 2020), and visual question answering (Wang et al. 2020b).

Depending on the solution process, existing REC methods can be broadly divided into one-stage and two-stage frameworks as shown in Fig. 1. The one-stage approaches (Sun, Xiao, and Lim 2021; Deng et al. 2021) treat the REC as an object detection with online classification, i.e., using referring expressions to define categories instead of a predefined set of categories. By extending object detectors, the conventional one-stage approaches (Sun, Xiao, and Lim 2021; Huang et al. 2021) utilize multi-head networks to model the comprehension (*cross-modal confidence*) and localization (*instance detection*) processes, cf., Fig. 1 (a). Inspired by *DETR* (Carion et al. 2020), transformer-based approaches (Deng et al. 2021) have recently received widespread attention as a flexible and effective framework. Leveraging the attention mechanism, these methods achieve deep *cross-modal alignment* and *query-based localization*, cf., Fig. 1 (b). One-stage methods, regardless of grounding or segmentation, are multi-objective implicitly coupled processes, i.e., the ability to comprehend abstract concepts is measured indirectly by their performance in the physical visual space (bounding box or segmentation mask). Although the one-stage framework achieves significant improvement by sufficiently exploiting the visual and linguistic context, these methods are limited by computation and data resources due to their complex optimization.

Two-stage approaches (Yu et al. 2018; Chen et al. 2021) attempt to build a matching and ranking process, which is a more natural scheme. As shown in Fig. 1 (c), conventional two-stage approaches usually merge the results of a pre-trained detector and cross-modal matching module at region-level, and search for the most relevant region proposal via a ranking process. Unfortunately, the conventional framework suffers from two inherent defects: 1) Sparse region proposals destroy the complete spatial context. 2) The ground-truth used during training and the proposals predicted by the detector form a gap, which leads to a sub-

*Corresponding author: Qingbao Huang

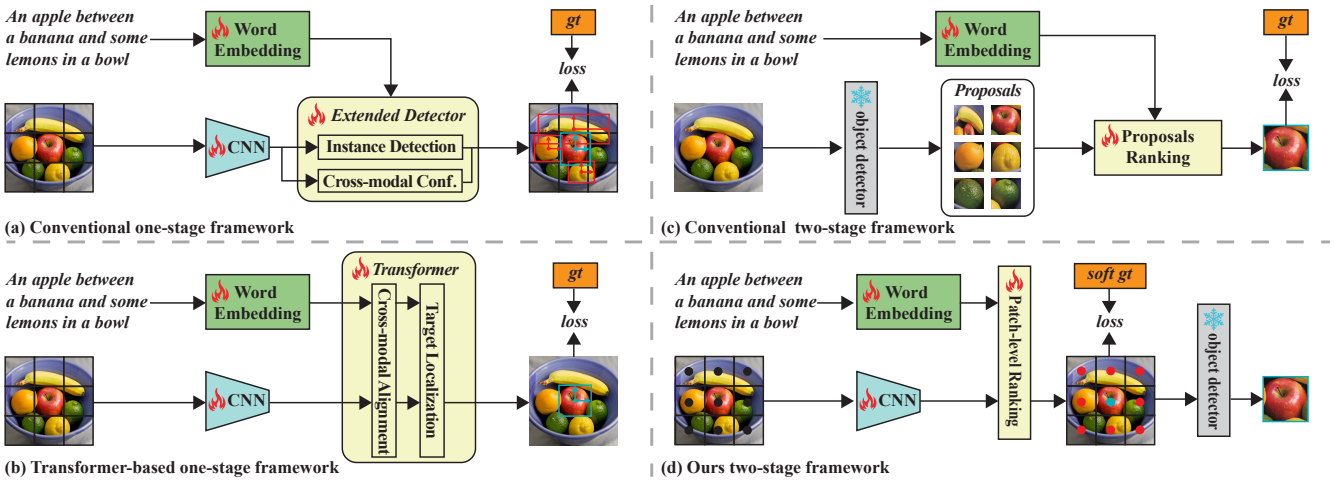

Figure 1: A comparison of (a) conventional one-stage framework, where *Conf.* means a confidence branch, (b) transformer-based one-stage framework, (c) conventional two-stage framework, and (d) our proposed point-based two-stage framework. Our method can leverage whole context information and merge comprehension and localization processes on the feature space.

optimal generalization of the model. In addition, the previous two-stage segmentation methods are a compromise implementation. The indirect solution for segmentation via the bounding box limits their performance.

In this paper, we propose a point-based two-stage framework for REC to address the aforementioned problems. As shown in Fig. 1 (d), instead of using the search space composed of loose and unordered regions, we propose to apply a regular set of points to support the ranking process. Specifically, relying on cross-modal fusion representations and point-based detectors (Tian et al. 2019), we reformulate the comprehension and localization processes of REC. To measure the correlation between visual points and referring expression, we construct soft ground-truth, e.g., center-ness and mass-ness matrices, based on the bounding boxes and segmentation masks. Then we establish a shape-independent classification process as the comprehension stage, which is an end-to-end trainable module optimized by the soft ground-truth. To convert the points from the comprehension stage into bounding boxes or segmentation masks, we introduce an IoU-based non-maximum suppression, which enables concise and efficient post-processing of predictions from the detector. Importantly, the shape-independent comprehension allows consistent modeling of grounding and segmentation tasks, so our framework supports multi-task learning without attaching any additional head network.

Our contributions can be summarized as: 1) We propose a point-based two-stage framework for REC. By approximating REC as a binary classification task, our framework can leverage the complete visual and linguistic context at a lower training cost. 2) We introduce soft ground-truth as the optimization objective of the cross-modal comprehension. Relying on the consistency of soft ground-truth in grounding and segmentation, our framework can naturally support multi-task learning, i.e., REG and RES. 3) Extensive experimental results on widely used benchmarks demonstrate the feasibility of the point-based paradigm. Our framework has significant improvements over conventional two-stage methods on both referring expression grounding and segmentation.

Related Work

Referring expression comprehension (REC) is originally described as retrieving a visual instance referred by a sentence from a set of region annotations. Thereby, early works (Yu et al. 2016) usually formulate the task as a ranking problem.

Two-stage inference frameworks replace the high-quality ground-truth for ranking with the region proposals of pre-trained object detectors, e.g., Faster-RCNN (Ren et al. 2015), to realize an automatic localization. However, most of the two-stage methods are motivated to reconstruct the context between regions because sparse proposals destroy the visual information. Module-based methods (Yu et al. 2018; Hu et al. 2017) decompose the alignment of multi-modal representations into several components. Yu et al. implicitly models the subject, relationship, and location by introducing different heuristic priors to different modules. Considering the multi-hop relationships between visual and linguistic instances, graph-based methods (Wang et al. 2019; Yang, Li, and Yu 2020; Sun et al. 2023) propose to construct regions and expressions as a scene graph or tree, which allows cross-modal representations to be aligned under the same structure. Two-stage frameworks reduce the training cost of REC by decoupling tasks. However, the incomplete visual semantics, especially discarded spatial context, hinder the alignment and fusion of multi-modal representations. Furthermore, the prediction proposals have a region shift compared to the high-quality annotations. This data discrepancy makes the ranking model struggle to generalize.

One-stage methods (Sadhu, Chen, and Nevatia 2019; Hu, Rohrbach, and Darrell 2016) recommend using an end-to-end process to solve REC, i.e., directly predicts the referred instance from the entire image and expression, which can eliminate the noise caused by the region proposals to the reasoning system. The conventional one-stage methods (Liao

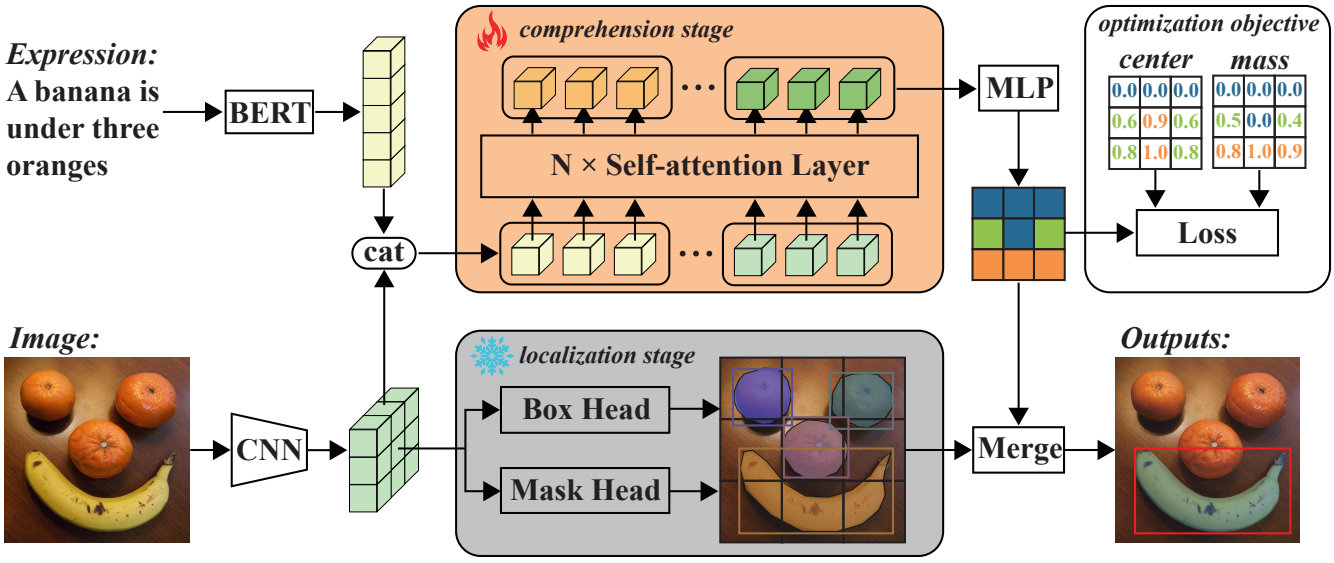

Figure 2: The overall framework of our model. It consists of two stages in inference: (1) a trainable cross-modal comprehension stage which is optimized by center-ness or mass-ness score, and (2) a frozen vision localization stage.

et al. 2020; Jing et al. 2021) generally implement grounding or segmentation by extending the natural language component to YOLOv3(Redmon and Farhadi 2018) or FCN (Long, Shelhamer, and Darrell 2015), respectively. For these cross-domain extended inference systems, effective fine-grained modeling is the key to localization. Yang et al. establish an iterative reasoning process for complex long expressions via sub-query. With the efficient cross-modal representations ability of the attention mechanism, Deng et al. proposed a transformer-based solution. Most recent works (Yang et al. 2022; Du et al. 2022; Zhu et al. 2022; Ye et al. 2022) have followed this new paradigm. Despite the significant performance improvement, these methods suffer from a long optimization process, usually requiring around 100 epochs. Therefore, the cost of data and computation is one of the most obvious limitations of transformer-based methods.

Approach

We present a point-based two-stage method for REC, a unified framework for the grounding and segmentation tasks based on the cross-modal comprehension and vision detection. As shown in Fig. 2, given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a referring expression $Q \in \mathbb{R}^{L}$, the task of REG is to predict the bounding box $\hat{b} \in \mathbb{R}^4$ of the referred instance, and the task of RES is to predict the segmentation mask $\hat{s} \in \mathbb{R}^{H\times W}$.

Problem Reformulation

The central idea of our point-based two-stage framework is to reformulate REC as an approximate binary classification problem. In conventional two-stage methods, anchor-based detectors, e.g., *Faster R-CNN* (Ren et al. 2015), cause a series of critical defects. The survey by Chen et al. (2021) shows that the recall of the region proposals obtained by the prevailing methods in inference is only 80.77%. In addition, previous methods usually crudely combine localization and comprehension at region-level. This makes it difficult for the comprehension module to generalize via the sub-optimal predictions. Inspired by FCOS (Tian et al. 2019), an object detector that projects points to bounding boxes in a one-to-one manner, we propose direct metrics of referring expression comprehension at point-level as soft ground-truth, e.g., expression-aware center-ness and mass-ness scores.

Expression-aware center-ness matrix is constructed by the ground-truth bounding box $b = (x_l, y_t, x_r, y_b)$, where (x_l, y_t) and (x_r, y_b) are the coordinates of the left-top and right-bottom corners. Concretely, we denote the feature maps extracted from the input image by visual backbone as $\boldsymbol{F}_v \in \mathbb{R}^{\frac{H}{g}\times\frac{W}{g}\times C}$, where g is the width of a visual grid. The k^{th} feature point in F_v represent the visual information collected from the grid (i, j), where $k = j \cdot \frac{W}{g} + i$. Similar to semantic segmentation, we assign categories to each point to indicate whether it is related to the referred instance. Since the bounding box belongs to low-precision localization, we define the points falling in the central area of the target box as positive samples to prevent noise from the irrelevant periphery. The central area is defined as a square box centered at $(\frac{x_l+x_r}{2}, \frac{y_t+y_b}{2})$ and the width is $3g$. According to the coordinates (x_k, y_k) of the k^{th} point projected on the raw image, the relative positional relationship between the positive point and the target bounding box can be defined as:

$$\begin{aligned} l &= x_k - x_l, \quad r = x_r - x_k, \\ t &= y_k - y_t, \quad b = y_b - y_k, \end{aligned} \tag{1}$$

where (l, r, t, b) is the left, right, top, and bottom distances from the point to the ground-truth bounding box. We compute the center-ness score by:

$$\boldsymbol{c}_k = \sqrt{\frac{\min(l, r)}{\max(l, r)} \cdot \frac{\min(t, b)}{\max(t, b)}}. \tag{2}$$

The center-ness score measures the degree how much the point deviates from the center of the target box, which allows the model to focus more on the grids near the center.

Expression-aware mass-ness matrix is roughly the same as center-ness, converted by the ground-truth segmentation mask $s \in \mathbb{R}^{H \times W}$, which is a boolean matrix used to segment the referred instance. Similarly, we define a score for each point on the feature map to measure its importance for prediction. As the segmentation mask is the best approximation to the instance shape at pixel-level, we assign positive samples to wider regions. Concretely, we first compute the centroid (x_m, y_m) of the target by the mask:

$$x_m = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad y_m = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{3}$$

where N is the total number of foreground pixels in the segmentation mask, and (x_i, y_i) is the pixel coordinate belonging to the foreground. Taking the centroid of the foreground as the center, all points falling in the area with a radius of $2g$ are considered as positive samples. Since the segmentation mask defines a binary class label for each pixel, we calculate the mass-ness as follows:

$$\boldsymbol{m}_k = \frac{1}{g^2}\sum_{x=1}^{g}\sum_{y=1}^{g} s_k(x, y), \tag{4}$$

where s_k is the k^{th} grid of the segmentation mask. We use m_k to quantify the foreground enrichment of each grid.

Both the center-ness and mass-ness of all negative points are set to 0. Using the above formulations, grounding and segmentation are approximated as a binary classification problem at the same scale, which makes the learning process of the model simpler and enables the same framework to solve multiple tasks.

Network Architecture

An overview of our model is shown in Fig. 2. The core design for the point-based two-stage framework is that two parallel reasoning stages are implemented by constructing soft ground-truth, i.e., point-based cross-modal comprehension and point-based localization.

For the feature encoding, given an image and a referring expression, we first extract the visual features F_v by a convolutional network (e.g., ResNet-101) and extract a sequence of textual tokens $\boldsymbol{F}_q \in \mathbb{R}^{L \times C}$ by BERT (Devlin et al. 2019). Then we utilize the point-based cross-modal comprehension module to align and fuse the multi-modal representations. As the key component in our model, the architecture of the comprehension stage is concise and elegant. For the uni-modal representations of $\boldsymbol{F}_v$ and $\boldsymbol{F}_q$, which are usually inconsistent in the channel dimension, we apply two linear layers to project them into the same embedding space. We denote the initial embedding as $\boldsymbol{F}_v^0$ and $\boldsymbol{F}_q^0$. Then we flatten and concatenate them as $\boldsymbol{F}_{vq}^0 = \{\boldsymbol{F}_v^0; \boldsymbol{F}_q^0\}$. To perform efficient intra- and inter-modal context interactions, we propose a visual-language transformer encoder that stacks a set of multi-head self-attention layers and feed-forward networks. The procedure in the encoder is formulated as:

$$\boldsymbol{F}_v' = \mathcal{F}_{trans}(\boldsymbol{F}_{vq}^0, \boldsymbol{e}_v)|_{0:g^2}, \tag{5}$$

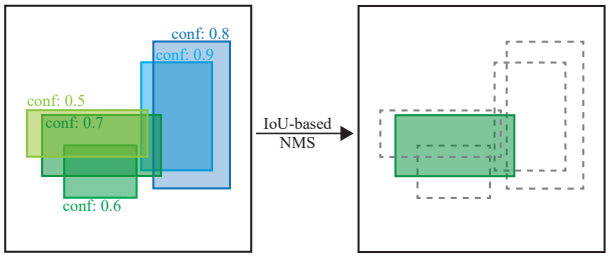

Figure 3: Our proposed IoU-based Non-Maximum Suppression (NMS), which can selects the final prediction by computing the IoU between a set of proposals.

where $\boldsymbol{e}_v \in \mathbb{R}^{\frac{HW}{g^2} \times C}$ is a flattened 2D-aware position embedding to compensate the absolute position information of the visual representation which is corrupted by convolutional translation invariance. We exploit the deep interactive visual state F_v' for classification prediction. Two shared MLP heads are utilized to obtain the center-ness prediction $\hat{\boldsymbol{c}}$ and mass-ness prediction $\hat{\boldsymbol{m}}$. Finally, the output of the prediction head is normalized by Sigmoid.

Consistent with conventional methods, our localization stage includes a generic object detection or segmentation model. However, the anchor-based methods, do not have a one-to-one mapping between predictions and feature points, which does not match our central idea. Therefore, we replace the box head and mask head with point-based models, e.g., FCOS (Tian et al. 2019) and SOLO (Wang et al. 2020a). We compare the ceiling of performance provided by different proposed methods in experiments.

Optimization and Inference

For the backward, we use expression-aware center-ness and mass-ness scores as the objectives to optimize our comprehension stage. As a multi-label binary classification task, our loss function is as follows:

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}) = \frac{1}{n^2}\sum_{i=1}^{n}(&\lambda_c \mathcal{L}_{BCE}(\boldsymbol{c}_i, \hat{\boldsymbol{c}}_i) \\ &+\lambda_m \mathcal{L}_{BCE}(\boldsymbol{m}_i, \hat{\boldsymbol{m}}_i)),\end{aligned} \tag{6}$$

where $n = \frac{HW}{g^2}$, which is the number of grids, $\mathcal{L}_{BCE}$ is binary cross entropy loss function, λ_c and λ_m are boolean values used to adjust the task type.

Relying on sparse proposals, conventional methods can take the top-1 region as the final prediction. However, our search space composed of points is a dense proposal set with a large amount of overlap. To take advantage of overlapping characteristics, we propose an IoU-based non-maximum suppression (NMS) as the post-process. As shown in Fig. 3, although expression-related regions tend to score higher, it is still possible that some high-confidence proposals are incorrect predictions. Therefore we recommend finding the maximum overlapping proposal as a reliable prediction. Specifically, for a set of proposals $\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k\}$, which is the top-k bounding box or segmentation mask ranked by

Detectors	RefCOCO			RefCOCO+			RefCOCOg	
	val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*test*
Faster R-CNN	98.25	99.40	98.45	98.38	99.41	98.77	97.39	97.22
Mask R-CNN	97.60	97.81	96.58	97.79	97.78	96.99	97.18	96.91
FCOS-P5	97.90	98.90	97.43	98.04	98.85	97.59	96.47	96.75
Mask R-CNN †	88.86	93.94	80.77	89.33	93.96	81.45	85.97	86.10

Table 1: Comparison of the recall (%) of different object detectors on RefCOCO, RefCOCO+, and RefCOCOg, †denotes the real case used in the state-of-the-art two-stage REC methods.

Models	Venue	Backbone	Epochs	RefCOCO			RefCOCO+			RefCOCOg	
				val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*test*
One-stage:											
FAOA	ICCV'2019	DarkNet-53	-	72.54	74.35	68.50	56.81	60.23	49.60	61.33	60.36
MCN	CVPR'2020	DarkNet-53	45	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
ReSC-Large	ECCV'2020	DarkNet-53	100	77.63	80.45	72.30	63.59	68.36	56.81	67.30	67.20
TransVG	ICCV'2021	ResNet-101	180	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
RefTR	NeurIPS'2021	ResNet-101	-	<u>82.23</u>	<u>85.59</u>	76.57	71.58	75.96	62.16	69.41	69.40
RED	AAAI'2022	DarkNet-53	100	80.97	83.20	77.66	69.48	73.80	62.20	71.11	70.67
Word2Pix	TNNLS'2022	ResNet-101	180	81.20	84.39	78.12	69.74	76.11	61.24	70.81	71.34
Two-stage:											
MAttNet	CVPR'2018	ResNet-101	5	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
CM-Att	CVPR'2019b	ResNet-101	5	78.35	83.14	71.32	68.09	73.65	58.03	67.99	68.67
Ref-NMS	AAAI'2021	ResNet-101	5	80.70	84.00	76.04	68.25	73.68	59.42	70.55	70.62
RvG-Tree	TPAMI'2022	ResNet-101	-	75.06	78.61	69.85	63.51	67.45	56.66	66.95	66.51
A-ATT	TPAMI'2022	VGG16	60	-	80.87	71.55	-	65.13	55.01	63.84	-
PBREC	ours	ResNet-101	15	82.20	85.26	<u>79.21</u>	<u>72.63</u>	<u>78.96</u>	<u>64.74</u>	**73.92**	<u>73.18</u>
PBREC-MT	ours	ResNet-101	15	**82.94**	**86.31**	**80.81**	**74.85**	**79.53**	**65.60**	<u>73.86</u>	**74.13**

Table 2: Comparison with the state-of-the-art REG approaches on RefCOCO, RefCOCO+, and RefCOCOg in terms of top-1 accuracy (%). The best and second best performances are in bold and underline, respectively.

center-ness or mass-ness, we compute the overlap score as:

$$r_i = \sum_{j=1}^{k} IoU(\hat{p}_i, \hat{p}_j). \tag{7}$$

Finally, we take the proposal with the highest overlap score as the final output. The detailed settings of IoU-based NMS are in the ablation experiments.

Experiments

Datasets and Evaluation Metrics

Follow Chen et al. (2021), we verify the effectiveness of our method on **RefCOCO** (Yu et al. 2016), **RefCOCO+** (Yu et al. 2016), and **RefCOCOg** (Mao et al. 2016). The images of these datasets are collected from MSCOCO (Lin et al. 2014). The three datasets have different challenges. The average sentence lengths of RefCOCO and RefCOCO+ are 3.50, 3.53, but RefCOCO+ prohibits the description of absolute positional relationships. RefCOCOg provides more realistic and complex expressions, and the average sentence length reaches 8.46. RefCOCOg has two types of splits, we use umd split which contains val and test set.

Following Deng et al. (2021), we evaluate REG by accuracy. When the IoU between the predicted bounding box and the ground truth is greater than 0.5, the prediction is deemed accurate. For the RES, we choose overall IoU as the metric which is obtained by computing the average IOU between the predicted mask and the ground-truth for all cases.

Implementation Details

We resize and pad all the images to 640×640, and follow Deng et al. (2021) to augment the raw data. We use ResNet-101 (He et al. 2016) as the vision backbone, and use the output of C5 block as the visual feature map, i.e., $g = 32$. For the tokenization, we set the max token length to 30 (RefCOCO, RefCOCO+, ReferItGame) and 40 (RefCOCOg). For the comprehension stage, we use a *6-layer* transformer encoder as our neck network. During training, we set batch size to 64, set the initial learning rate to 1×10^{-4} for comprehension module, set a lower initial learning rate 1×10^{-6} for ResNet and BERT. The model is dynamically optimized for **15** epochs by AdamW (Loshchilov and Hutter 2019) and CosineAnnealing (Loshchilov and Hutter 2017). During inference, we take the P5 block of FCOS (Tian et al. 2019) and P4 block of SOLO (Wang et al. 2020a) as the grounding and segmentation proposal source, and set k to 12 in the IoU-based NMS. We provide two versions of the model, i.e., PBREC optimized for single task and PBREC-MT optimized for multi-task.

Models	Venue	Backbone	RefCOCO			RefCOCO+			RefCOCOg	
			val	*testA*	*testB*	*val*	*testA*	*testB*	*val*	*test*
One-stage:										
LTS	CVPR'2021	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT	ICCV'2021	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
RefTR	NeurIPS'2021	ResNet-101	70.56	**73.49**	66.57	61.08	64.69	52.73	58.73	58.51
ResTR	CVPR'2022	ViT-16	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-
CRIS	CVPR'2022	ResNet-101	70.47	73.18	66.10	62.27	**68.08**	53.68	59.87	60.36
SeqTR	ECCV'2022	DarkNet-53	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
Two-stage:										
MAttNet	CVPR'2018	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree	ICCV'2019a	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CM-Att	CVPR'2019b	ResNet-101	58.23	64.60	53.14	49.65	53.90	41.77	49.10	50.72
Ref-NMS	AAAI'2021	ResNet-101	61.46	65.55	57.41	49.76	53.84	42.66	51.21	51.90
PBREC	ours	ResNet-101	<u>71.11</u>	72.89	**70.12**	<u>62.99</u>	66.67	**56.64**	<u>62.14</u>	<u>61.56</u>
PBREC-MT	ours	ResNet-101	**71.44**	<u>73.21</u>	<u>70.11</u>	**63.76**	<u>67.10</u>	<u>56.63</u>	**62.93**	**62.61**

Table 3: Comparison with the state-of-the-art RES approaches on RefCOCO, RefCOCO+, and RefCOCOg in terms of overall IoU (%). The best and second best performance are in bold and underline, respectively.

	Grounding		Segmentation	
	hard	*soft*	*hard*	*soft*
Single task	71.93	-	61.20	-
	-	72.63	-	62.99
Multi task	72.17	-	62.25	-
	71.96	-	-	62.45
	-	72.75	62.04	-
	-	**74.85**	-	**63.76**

Table 4: Ablation study of center-ness and mass-ness.

Top-k	REG		RES
	center	center+conf	mass
1	73.55	73.24	63.34
4	74.14	74.13	64.07
8	74.33	74.46	**64.08**
12	**74.51**	**74.85**	63.76
16	74.25	74.64	62.23

Table 5: Ablation study of IoU-based NMS.

Comparison with State-of-the-art Models

Compared with prevailing two-stage methods, one notable difference is that we use a different pre-trained detector. To make a more convincing performance comparison, we follow the statistics of Chen et al. (2021), to compare the ceiling of performance that different detectors can provide. As shown in Table 1, we compute the recall of region proposals in several different scenarios, i.e., the proportion that the proposals contain correct prediction. We have the following observations: 1) When using the top-100 GreedyNMS (cf. Row 1 and Row 2), which is a usual practice for most downstream tasks, the recall of anchor-based detectors can reach about 97%. 2) The predictions at P5 level using FCOS improve the performance ceiling by less than 1%. This is reasonable, since we provide more proposals (typically 400 boxes). 3) To reduce the gap between training and inference, prevailing two-stage methods usually use sparse proposals (e.g., less than 10) in the real case. This is an obvious performance bottleneck, e.g., it is impossible for these methods on RefCOCO testB to exceed 80.77%.

To evaluate our method, we compare it with other state-of-the-art methods on grounding and segmentation tasks. The REG performance is shown in Table 2. Compared with the emerging conventional one-stage method RED (Huang et al. 2022), our model obtains absolute improvements by 1.97%-3.15%, 3.40%-5.73%, and 2.81%-3.46% on RefCOCO, RefCOCO+, and RefCOCOg, respectively. When comparing to TransVG (Deng et al. 2021), a transformer-based method most similar to our neck architecture, our method requires shorter training epochs (15 vs 180) to achieve better performance, with 3.59%/ 10.03%/ 6.40% on RefCOCO (testA), RefCOCO+ (val), and RefCOCOg (test), respectively. That means our approximate classification greatly reduces the learning difficulty of the task and achieves significant performance improvement by relying on task decoupling. Our model also achieves obvious improvements compared with all two-stage methods. Specifically, our model outperforms the recent state-of-the-art method Ref-NMS (Chen et al. 2021) by 4.77%, 6.60%, and 3.51% on the three datasets. Notably, on the RefCOCO(testB), the limit performance of the conventional method is 80.77% (cf. Row 4 of Table 1), while our method can reach 80.81%.

For the RES task, we summarize the performance comparison in Table 3. Compared with two-stage methods, our model has an absolute advantage with 9.98%/ 7.66%/ 12.70%, 14.00%/ 13.26%/ 13.98%, and 11.72%/ 10.71% on RefCOCO, RefCOCO+, and RefCOCOg, respectively. The significant performance gap shows that previous segmentation methods are limited by the bounding box, while our mass-ness metric is a reasonable solution. Our method is also competitive with one-stage methods. Compared with

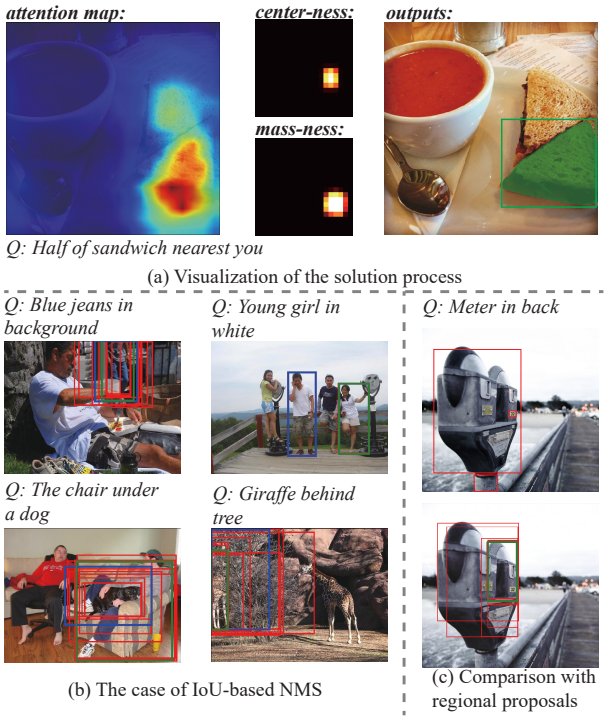

Figure 4: Visualization of cases,the red box is the top-k proposals, the blue box is the top-1 prediction, and the green box is the correct prediction by our model.

CRIS (Wang et al. 2022), which is a CLIP-based knowledge transfer model, our performance improves at most 4.02%.

Ablation Study

We use PBREC-MT to conduct ablation studies, and all performance changes are verified on RefCOCO+ (val).

Soft ground-truth: To verify the rationality of our designed center-ness and mass-ness scores, we try a number of different combinations of metrics. Table 4 shows the performance changes of the model in a single task training or multi-task joint training, where *hard* means that the classification task is directly modeled with 0-1, and *soft* means that the category is represented by our designed scores. Compared with hard classification, our metrics improve grounding and segmentation performance by +0.70%/ +1.79% and +2.68%/ +1.51% in single-task and multi-task, respectively. Furthermore, we observe a noteworthy phenomenon in multi-task optimization. Concretely, for the baseline performance 72.17%/ 62.25%, using only center-ness brings a change of +0.58%/ -0.21%, and mass-ness brings a change of -0.21%/ +0.20%. In this case, soft ground-truth not only fails to bring significant improvement, but also leads to another task performance penalty. This means the right combination is even more important for multi-task learning.

IoU-based NMS: To find an appropriate post-processing setting, we compare the performance of several different schemes. As shown in Table 5, for the REG task, since FCOS also uses the degree of center deviation to measure the prediction confidence, we try two ranking schemes, i.e., using center-ness alone and using the product of center-ness and FCOS confidence. The SOLO model uses the quality of the prediction mask as confidence, which does not fit our motivation, so we only use mass-ness as the ranking basis. Similar to conventional methods, direct top-1 ranking can provide considerable performance. There is a further increase in performance when performing IoU-based NMS on the top-k predictions. Taking REG as an example, the improvement reaches maximum at top-12 (+1.61%), but further expansion of the group will lead to performance degradation. This improvement is not obvious on RES because the overall IoU is a pixel-level metric and is not easily affected by noise. Finally, without loss of generality, we use top-12 post-processing as the unified setting.

Qualitative Results

We illustrate the qualitative results with case visualizations. As shown in Fig. 4 (a), we visualize the inference process. The attention map is taken from the scores of the textual [CLS] token and all visual grids in the last layer of the comprehension stage. It can be observed that our model comprehends the abstract concept of sandwich, and more focus on the nearest sandwich in the image. Both the center-ness and mass-ness score predictions are successfully focused on the referred instance. The visualization in Fig. 4 (b) shows the effect of our IoU-based NMS from four cases: 1) The left top one shows that the noise generated by occlusion causes the top-1 prediction area to become larger, while our NMS can refine the predictions; 2) the left bottom one shows another refinement process, i.e., the prediction box is extended outward; 3) the right top one shows that a few parts of the box only perceives *'white'* but not the *'girl'*, which leads to an incomplete cross-modal comprehension, however, our NMS is a majority filter, which can correct this error; 4) the right bottom one demonstrates the ability of our NMS to mine samples that are difficult to locate. The case in Fig. 4 (c) verifies a key conclusion. Specifically, this case is to localize a *'meter'* which is behind the other. For conventional methods (top), the first stage can only provide three expression-independent boxes as proposals when using 0.65 as the confidence threshold. In fact, the target is included in proposals with even lower confidence thresholds. According to the idea of doing comprehension first and then localizing, our method (bottom) uses the expression-aware metrics as the basis for ranking, leading to more accurate localization.

Conclusions

In this paper, we propose a novel two-stage REC paradigm, which achieves a point-based modulate localization by approximating grounding or segmentation as a classification problem. With the parallel inference framework and point-level metrics, e.g., center-ness and mass-ness, we overcome the inherent defects of prevailing two-stage methods, thereby breaking through performance bottlenecks. Extensive experiments demonstrate the feasibility of our method. In the future, we plan to develop our point-based two-stage paradigm in the open domain, zero-shot, etc.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62276072), the Guangxi Natural Science Foundation (No. 2022GXNSFAA035627), National Natural Science Foundation of China (62076100), Guangxi Scientific and Technological Bases and Talents Special Projects (guikeAD23026230 and guikeAD23026213), Guangxi Natural Science Foundation Key Project (Application No. 2023JJD170015), the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, the Fundamental Research Funds for the Central Universities, SCUT (D2230080), Innovation Project of Guangxi Graduate Education, CAAI-Huawei MindSpore Open Fund and the Science and Technology Planning Project of Guangdong Province (2020B0101100002).

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12346 of *Lecture Notes in Computer Science*, 213–229.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S. 2021. Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 1036–1044.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 9959–9968.

Deng, C.; Wu, Q.; Wu, Q.; Hu, F.; Lyu, F.; and Tan, M. 2022. Visual Grounding Via Accumulated Attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3): 1670–1684.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 1749–1759.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 4171–4186.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 16301–16310.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Visual Grounding with Transformers. In *IEEE International Conference on Multimedia and Expo, ICME 2022*, 1–6.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 770–778.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2022. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2): 684–696.

Hu, R.; Rohrbach, M.; Andreas, J.; Darrell, T.; and Saenko, K. 2017. Modeling Relationships in Referential Expressions with Compositional Modular Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 4418–4427.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from Natural Language Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9905 of *Lecture Notes in Computer Science*, 108–124.

Huang, B.; Lian, D.; Luo, W.; and Gao, S. 2021. Look Before You Leap: Learning Landmark Features for One-Stage Visual Grounding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual*, 16888–16897.

Huang, J.; Qin, Y.; Qi, J.; Sun, Q.; and Zhang, H. 2022. Deconfounded Visual Grounding. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, 998–1006.

Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; and Tan, T. 2021. Locate Then Segment: A Strong Pipeline for Referring Image Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 9858–9867.

Kim, J.; Misu, T.; Chen, Y.; Tawari, A.; and Canny, J. F. 2019. Grounding Human-To-Vehicle Advice for Self-Driving Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 10591–10599.

Kim, N.; Kim, D.; Kwak, S.; Lan, C.; and Zeng, W. 2022. ReSTR: Convolution-free Referring Image Segmentation Using Transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 18124–18133. IEEE.

Li, M.; and Sigal, L. 2021. Referring Transformer: A One-step Approach to Multi-task Visual Grounding. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 19652–19664.

Liao, Y.; Liu, S.; Li, G.; Wang, F.; Chen, Y.; Qian, C.; and Li, B. 2020. A Real-Time Cross-Modality Correlation Filtering Method for Referring Expression Comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 10877–10886.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, 740–755.

Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; and Yuille, A. L. 2017. Recurrent Multimodal Interaction for Referring Image Segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017*, 1280–1289.

Liu, D.; Zhang, H.; Zha, Z.; and Wu, F. 2019a. Learning to Assemble Neural Module Tree Networks for Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 4672–4681.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019b. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 1950–1959.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 10031–10040.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 11–20.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement. *CoRR*, abs/1804.02767.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 91–99.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-Shot Grounding of Objects From Natural Language Queries. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 4693–4702.

Sun, M.; Xiao, J.; and Lim, E. G. 2021. Iterative Shrinking for Referring Expression Grounding Using Deep Reinforcement Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 14060–14069.

Sun, M.; Xiao, J.; Lim, E. G.; and Zhao, Y. 2023. Cycle-Free Weakly Referring Expression Grounding With Self-Paced Learning. *IEEE Trans. Multim.*, 25: 1611–1621.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9626–9635. IEEE.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2019. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 1960–1968.

Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; and Li, L. 2020a. SOLO: Segmenting Objects by Locations. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, 649–665. Springer.

Wang, X.; Liu, Y.; Shen, C.; Ng, C. C.; Luo, C.; Jin, L.; Chan, C. S.; van den Hengel, A.; and Wang, L. 2020b. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 10123–10132.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. CRIS: CLIP-Driven Referring Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 11676–11685. IEEE.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 9489–9498.

Yang, S.; Li, G.; and Yu, Y. 2020. Graph-Structured Referring Expression Reasoning in the Wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 9949–9958.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving One-Stage Visual Grounding by Recursive Sub-query Construction. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12359, 387–404.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A Fast and Accurate One-Stage Approach to Visual Grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 4682–4692.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 15481–15491.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference*, volume 9906 of *Lecture Notes in Computer Science*, 69–85.

Zhao, H.; Zhou, J. T.; and Ong, Y.-S. 2022. Word2Pix: Word to Pixel Cross-Attention Transformer in Visual Grounding. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.

Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. SeqTR: A Simple Yet Universal Network for Visual Grounding. In *Computer Vision - ECCV 2022 - 17th European Conference*, volume 13695 of *Lecture Notes in Computer Science*, 598–615.