SeqTR: A Simple yet Universal Network for Visual Grounding

Chaoyang Zhu[1], Yiyi Zhou[1], Yunhang Shen[3], Gen Luo[1], Xingjia Pan[3], Mingbao Lin[3], Chao Chen[3], Liujuan Cao[1*], Xiaoshuai Sun[1,4], Rongrong Ji[1,2,4]

[1]MAC Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University. [2]Institute of Energy Research, Jiangxi Academy of Sciences. [3]Tencent Youtu Lab. [4]Institute of Artificial Intelligence, Xiamen University.
cyzhu@stu.xmu.edu.cn, zhouyiyi@xmu.edu.cn, shenyunhang01@gmail.com, luogen@stu.xmu.edu.cn, xjia.pan@gmail.com, linmb001@outlook.com, aaronccchen@tencent.com, {caoliujuan,xssun,rrji}@xmu.edu.cn

Abstract. In this paper, we propose a simple yet universal network termed *SeqTR* for visual grounding tasks, *e.g.*, phrase localization, referring expression comprehension (REC) and segmentation (RES). The canonical paradigms for visual grounding often require substantial expertise in designing network architectures and loss functions, making them hard to generalize across tasks. To simplify and unify the modeling, we cast visual grounding as a point prediction problem conditioned on image and text inputs, where either the bounding box or binary mask is represented as a sequence of discrete coordinate tokens. Under this paradigm, visual grounding tasks are unified in our SeqTR network without task-specific branches or heads, *e.g.*, the convolutional mask decoder for RES, which greatly reduces the complexity of multi-task modeling. In addition, SeqTR also shares the same optimization objective for all tasks with a simple *cross-entropy* loss, further reducing the complexity of deploying hand-crafted loss functions. Experiments on five benchmark datasets demonstrate that the proposed SeqTR outperforms (or is on par with) the existing state-of-the-arts, proving that a simple yet universal approach for visual grounding is indeed feasible. Source code is available at https://github.com/sean-zhuh/SeqTR.

Keywords: Visual Grounding, Transformer

1 Introduction

Visual grounding [55,34,36,21,52] has emerged as a core problem in vision-language research, as both comprehensive intra-modality understanding and accurate one-to-one inter-modality correspondence establishment are required. According to the manner of grounding, it can be divided into two groups, *i.e.*, *phrase localization* or *referring expression comprehension* (REC) at bounding

*Corresponding author.

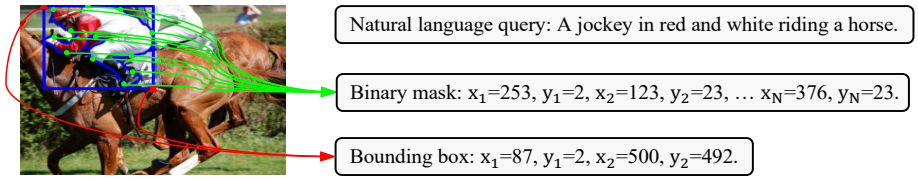

Fig. 1. Illustration of the serialization of grounding information. Our model directly generates the sequence of points representing the bounding box or binary mask.

box level [54,29,27,58,50,25,32,49,44,16,6,59,24], and *referring expression segmentation* (RES) at pixel level [54,51,2,15,18,28,32,10,48,31,19,8,24].

To accomplish the accurate vision-language alignment, existing approaches often require substantial prior knowledge and expertise in designing network architectures and loss functions. For instance, MAttNet [54] decomposes language expressions into *subject*, *location*, and *relationship* phrases, and designs three corresponding attention modules to compute matching score individually. Despite being faster, one-stage models also require the complex language-guided multi-modal fusion and reasoning modules [33,16,25,49,32], or sophisticated cross-modal alignment via various attention mechanisms [10,31,51,15,28,32,8]. Loss functions in existing methods are also complex and tailored to each individual grounding task, such as GIoU loss [41], set-based matching loss [1], focal loss [26], dice loss [35], and contrastive alignment loss [20]. Under a multi-task setting, coefficients among different losses also need to be carefully tuned to accommodate different tasks [32,24]. Despite great progress, these highly customized approaches still suffer from the limited generalization ability.

Recent endeavors [6,8,20,24] in visual grounding shift to simplifying network architectures via Transformers [45]. Concretely, the multi-modal fusion and reasoning modules are replaced by a simple stack of transformer encoder layers [6,20,8]. However, the loss function used in these transformer-based methods is still highly customized for each individual task [26,35,20,41,1]. Moreover, these approaches still require task-specific branches or heads [32,24], *i.e.*, the bounding box regressor and convolutional mask decoder.

In this paper, we take a step forward in simplifying the modeling of visual grounding tasks via a simple yet universal network termed *SeqTR*. Specifically, inspired by the recently proposed Pix2Seq [3], we first reformulate visual grounding as a point prediction problem conditioned on image and text inputs, where the grounding information, *e.g.*, the bounding box, is serialized into a sequence of discrete coordinate tokens. Under this paradigm, different grounding tasks can be universally accomplished in the proposed SeqTR with a standard transformer encoder-decoder architecture [45]. In SeqTR, the encoder serves to update the multi-modal feature representations, while the decoder directly predicts the discrete coordinate tokens of the grounding information in an auto-regressive manner. In terms of optimization, SeqTR only uses a simple *cross-entropy* loss for all grounding tasks, requiring no further prior knowledge or expertise. Over-

all, the proposed SeqTR greatly reduces the difficulty and complexity of both architecture design and optimization for visual grounding.

Notably, the proposed SeqTR is not just a simple multi-modal extension of Pix2Seq for the challenging open-ended visual grounding tasks. In addition to bridging the gap between object detection and visual grounding, we also apply the sequential modeling to RES via an innovative *mask contour sampling* scheme. As shown in Fig. 1, SeqTR transforms the pixel-wise binary mask into a sequence of N points by performing clockwise sampling on the mask contour. In this case, RES, as a language-guided segmentation task, can be seamlessly integrated into the proposed SeqTR network without the additional convolutional mask decoder, demonstrating the high generalization ability of SeqTR across grounding tasks.

The proposed SeqTR achieves or is on par with the state-of-the-art performance on five benchmark datasets, *i.e.*, RefCOCO [55], RefCOCO+ [55], RefCOCOg [34,36], ReferItGame [21], and Flickr30K Entities [37]. SeqTR also outperforms a set of large-scale BERT-style models [30,43,4,20] with much less pre-training expenditure. Main contributions are summarized as follows:

- We reformulate visual grounding tasks as a point prediction problem, and present a novel and general network, termed SeqTR, which unifies different grounding tasks in one model with the same *cross-entropy* loss.
- The proposed SeqTR is simple yet universal, and can be seamlessly extended to the referring expression segmentation task via an innovative *mask contour sampling* scheme without network architecture modifications.
- We achieve or maintain on par with the state-of-the-art performance on five visual grounding benchmark datasets, and also outperform a set of large-scale pre-trained models with much less expenditure.

2 Related Work

2.1 Referring Expression Comprehension

Early practitioners [14,57,60,54,29,47,13,27] tackle referring expression comprehension (REC) following a two-stage pipeline, where region proposals [40] are first extracted then ranked according to their similarity scores with the language query. Another line of work [58,50,25,32,49,44,16,6,59], being simpler and faster, advocates one-stage pipeline based on dense anchors [40]. RealGIN [58] proposes adaptive feature selection and global attentive reasoning unit to handle the diversity and complexity of language expressions. ReSC [49] recursively constructs sub-queries to predict the parameters of the normalization layers in the visual encoder, which is used to scale and shift visual features. LBYL [16] designs landmark feature convolution to encode the contextual information. Recent works [6,59,20,8,24] resort to Transformer-like structure [45] to perform multi-modal fusion. MDETR [20] further demonstrates that Transformer is efficient when pre-trained on a large corpus of data. Compared with existing approaches, our work is simple in both the architecture and loss function, which has little requirement of task priors and expert engineering.

2.2 Referring Expression Segmentation

Compared to REC, referring expression segmentation (RES) grounds language query at a fine-granularity *i.e.*, the precise pixel-wise binary mask. Typical solutions are to design various attention mechanisms to perform cross-modal alignment [31,32,8,51,10,2,15,17,28,18]. EFN [10] transforms the visual encoder into a multi-modal feature extractor with asymmetric co-attention, which fuses multi-modal information at the feature learning stage. CGAN [31] performs cascaded attention reasoning with instance-level attention loss to supervise attention modeling at each stage. LTS [19] first performs relavance filtering to locate the referent, and uses this visual object prior to perform dilated convolution for the final segmentation mask. VLT [8] produces a set of queries representing different understandings of the language expression and proposes a query balance module to focus on the most reasonable and suitable query, which is then used to decode the mask via a mask decoder. In this work, we are the first to regard RES as a point prediction problem, thus the proposed SeqTR can be seamlessly extended to RES without any network architecture modifications.

2.3 Multi-task Visual Grounding

Multi-task visual grounding aims to jointly address REC and RES. Prior art MCN [32] constrains the REC and RES branches to attend to the same region by applying consistent energy maximization. In this way, REC can help RES better localize the referent, and RES can help REC achieve superior cross-modal alignment. RefTR [24] tackles multi-task visual grounding by sharing the same transformer architecture, but it requires an additional convolutional mask decoder for RES. In contrast, the proposed SeqTR is universal across different grounding tasks without additional branch or head. Under the point prediction paradigm, SeqTR can segment the referent without the aid from REC branch.

3 Method

In this section, we introduce our simple yet universal SeqTR network for visual grounding, of which structure is depicted in Fig. 2. The objective function is detailed in Sec. 3.1. Sequence construction from grounding information is elaborated in Sec. 3.2. The architecture and inference are presented in Sec. 3.3.

3.1 Problem Definition

Unlike existing visual grounding models [32,6,10,19,8], SeqTR aims to predict the discrete coordinate tokens of the grounding information, *e.g.*, the bounding box or binary mask. To this end, we define the optimization objective under the point prediction paradigm as:

$$\mathcal{L} = -\sum_{i=1}^{2N} w_i \log P(T_i | F_m, S_{1:i-1}), \tag{1}$$

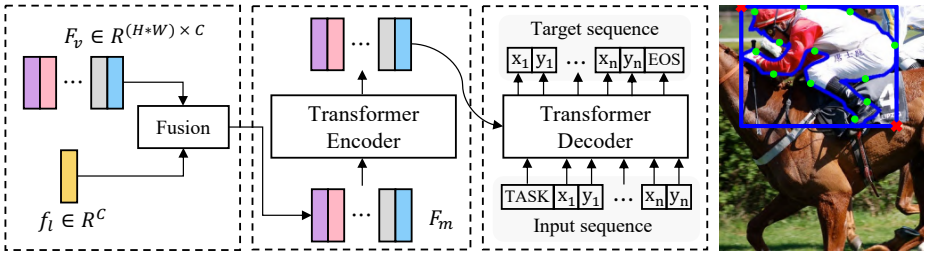

Fig. 2. Overview of the proposed SeqTR network, of which all components, *i.e.*, multi-modal fusion, cross-modal interaction, and loss function, are standard operations and shared across grounding tasks.

where S and T are the input and target sequences for decoder as shown in Fig. 2. $F_m \in R^{(H*W)\times C}$ is the multi-modal features detailed in Sec. 3.3. A per-token weight w_i is used to scale the loss. Note that the input sequence $S_{1:i-1}$ only contains the preceding coordinate tokens when predicting the i-th one. It can be implemented by putting a causal mask [38] on attention weights to only attend to previous coordinate tokens.

We construct the input sequence by prepending a [TASK] token before the sequence of points $\{x_i, y_i\}_{i=1}^N$, and the target sequence is the one appended with an [EOS] token. These two special tokens indicate the start or end of the sequence, which are learnable embeddings. [TASK] token also indicates which grounding task the model performs on. To achieve multi-task visual grounding, we can equip each task with the corresponding [TASK] token randomly initialized with different parameters, showing great simplicity and generalization ability.

Under our point prediction reformulation, the simple *cross-entropy* loss conditioned on multi-modal features and preceding discrete coordinate tokens can be directly shared across tasks, avoiding the complex deployment of hand-crafted loss functions and loss coefficient tuning [35,26,1,20,41].

3.2 Sequence Construction from Grounding Information

A key design in SeqTR is to serialize and quantize the grounding information, *e.g.*, the bounding box or binary mask, into a sequence of discrete coordinate tokens, which enables different grounding tasks to be universally addressed in one network architecture with the same objective.

We first review the serialization and quantization of the bounding box introduced in Pix2Seq [3]. Given a sequence of floating points $\{\tilde{x}_i, \tilde{y}_i\}_{i=1}^N$ representing the top-left and bottom-right corner points of the bounding box (N is 2), these floating coordinates are quantized into integer bins by

$$x_i = \text{round}(\frac{\tilde{x}_i}{w} * M), \qquad y_i = \text{round}(\frac{\tilde{y}_i}{h} * M), \tag{2}$$

where each coordinate is normalized by image width w and height h, and M is the number of quantization bins. We refer readers to Pix2Seq [3] for more

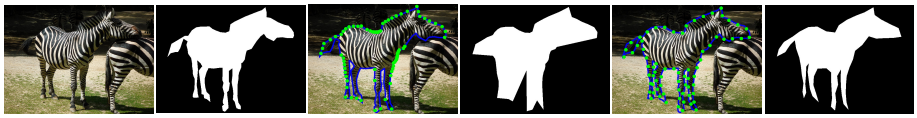

Fig. 3. Visualization of different sampling strategies. (a-b) are the original image and ground-truth. (c-d) are the sampled points and reassembled mask of center-based sampling, respectively, while (e-f) are the ones of uniform sampling.

discretization details. In practice, we construct a shared embedding vocabulary $E \in R^{M \times C}$ for both x-axis and y-axis.

While bounding boxes can be naturally determined by two of its corner points and serialized into a sequence as in Eq. 2, binary masks can not. A binary mask consists of infinite points, of which both quantities and positions impact the details of the mask significantly, thus the above serialization and quantization for bounding boxes is not directly applicable to binary masks.

To address this issue, we propose an innovative *mask contour sampling* scheme for the sequence construction from binary masks. As shown in Fig. 3, we sample N points clockwise from the consecutive mask contour of the referred object, then, the sequence of sampled points can be quantized via Eq. 2. Following sampling strategies are experimented:

- **Center-based sampling**. Starting from the mass center of the binary mask, N rays are emitted with the same angle interval. The intersection points between these rays and the mask contour are clockwise sampled.
- **Uniform sampling**. We uniformly sample N points clockwise on top of the mask contour, which is much simpler compared to the first strategy.

Compared to the center-based sampling, uniform sampling distributes the sampled points along the mask contour more evenly, and can better represent the irregular mask especially when the outline between two adjacent sampled points is tortuous. As shown in Fig. 3, center-based sampling loses the fine details of the zebra legs, while uniform sampling preserves the mask contour more precisely.

In practice, the proposed sampling scheme slightly restricts the performance upper-bound of RES, *e.g.*, uniformly sampling 36 points from ground-truth masks will achieve 95.63 mIoU on RefCOCO *validation* set. Considering current state-of-the-art performance, such a defect is still acceptable. Besides, even if we take as ground-truth the precise binary mask, the upper-bound still will not reach 100 mIoU since down-sampling operations are often necessary.

Both center-based and uniform sampling use deterministic (clockwise) ordering in the sequence of points for the binary mask, however, a binary mask is only determined by points' positions instead of the ordering. Hence we randomly shuffle points' order, which enables the model to learn which point to predict next. In Sec. 4.5, we thoroughly study the proposed sampling scheme.

3.3 Architecture

Language Encoder. To demonstrate the efficacy of SeqTR, we do not opt for the pre-trained language encoders such as BERT [7], hereby the language encoder is a one layer bidirectional GRU [5]. We concatenate both unidirectional hidden states $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$ at each step t to form word features $\{h_t\}_{t=1}^{T}$.

Visual Encoder. The multi-scale features of the visual encoder are unidirectionally down-sampled from the finest to coarsest spatial resolution, and flattened to generate visual features $F_v \in R^{(H*W)\times C}$ as input to the fusion module. H and W are 32 times smaller of the original image size. In contrast to previous work, we only use the coarsest scale visual features instead of the finest ones for RES task [32,19,10,31], as we do not predict the binary mask pixel-wisely, which reduces the memory footprint during training.

Fusion. Different from Pix2Seq [3], which only perceives the pixel inputs, we devise a simple yet efficient fusion module to align vision and language modalities. Given visual features F_v and word features $\{h_t\}_{t=1}^{T}$, we first construct language feature $f_l \in R^C$ by *max* pooling word features along the channel dimension. We use Hadamard product between F_v and f_l without the linear projection to produce the multi-modal features $F_m \in R^{(H*W)\times C}$ to transformer encoder:

$$F_{m,i} = \sigma(F_{v,i}) \odot \sigma(f_l), \tag{3}$$

where σ is tanh function. Note that we do not concatenate word features and visual features then use the transformer encoder to perform fusion as in [6,20,8], because that the complexity will quadratically increase.

Transformer and Predictor. The standard transformer encoder updates the feature representations of multi-modal features F_m, while the decoder predicts the target sequence in an auto-regressive manner. The hidden dimension of transformer is set to 256, the expansion rate in feed forward network (FFN) is 4, and the number of encoder and decoder layers are 6 and 3, respectively. This results in the transformer being extremely compact. Since the transformer is permutation-invariant, the F_m and the input sequence are added with *sine* and *learned* positional encoding [45], respectively. To predict the coordinate tokens, an MLP with a final softmax function is used.

Inference. During inference, coordinates are generated in an auto-regressive manner, each coordinate is the *argmax*-ed index of the probabilities over the vocabulary E, and mapped back to the original image scale via the inversion of Eq. 2. We predict exactly 4 discrete coordinate tokens for REC, while leaving the decision of when to prediction to [EOS] token for RES. The predicted sequence is assembled to form the bounding box or binary mask for evaluation.

4 Experiments

4.1 Datasets

RefCOCO/RefCOCO+/RefCOCOg. RefCOCO [55] contains 142,210 referring expressions, 50,000 referred objects, and 19,994 images. Referring expressions in testA set mostly describe people, while the ones in testB set mainly

describe objects except people. Similarly, RefCOCO+ [55] contains 141,564 expressions, 49,856 referred objects, and 19,992 images. Compared to RefCOCO, referring expressions of RefCOCO+ describe more about attributes of the referent, *e.g.*, color, shape, digits, and avoid using words of absolute spatial location. RefCOCOg [34,36] has two types of partition strategy, *i.e.*, the *google* split [34] and *umd* split [36]. Both splits have 95,010 referring expressions, 49,822 referred objects, and 25,799 images. We use the validation set as the test set following [10,48,16,49] for *umd* split. The language length of RefCOCOg is 8.4 words on average while that of RefCOCO and RefCOCO+ are only 3.6 and 3.5 words.
ReferItGame [21] contains 120,072 referring expressions and 99,220 referents for 19,997 images collected from the SAIAPR-12 [9] dataset. We use the cleaned berkeley split to partition the dataset, which consists of 54,127, 5,842, and 60,103 referring expressions in train, validation, and test set, respectively.
Flickr30K. Language queries in Flickr30K Entities [37] are short region phrases instead of sentences which may contain multiple objects. It contains 31,783 images with 427K referred entities in train, validation, and test set.
Pre-training dataset. Following [20], we merge region descriptions from Visual Genome (VG) [23] dataset, annotations from RefCOCO [55], RefCOCO+ [55], RefCOCOg [34,36], and ReferItGame [21] datasets, and Flickr entities [37]. This results in approximately 6.1M distinct language expressions and 174k images in train set, which are less than 200k images as in [20].

4.2 Evaluation Metrics

For REC and phrase localization, we evaluate the performance using Precision@0.5. The prediction is deemed correct if its intersection over union (IoU) with ground-truth box is larger than 0.5. For RES, we use $mIoU$ as the evaluation metric. Precision at 0.5, 0.7, and 0.9 thresholds are also used for ablation.

4.3 Implementation Details

We train SeqTR 60 epochs for REC and phrase localization, and 90 epochs for RES with batch size 128. The Adam [22] optimizer with an initial learning rate 5e-4 is used, which decays the learning rate 10 times after 50 epochs and 75 epochs for the detection and segmentation grounding tasks, respectively. Following standard practices [6,32,19,8], image size is resized to 640×640, and the length of language expression is trimmed at 15 for RefCOCO/+ and 20 for RefCOCOg. For ablation, we train SeqTR 30 epochs unless otherwise stated. During pre-training, SeqTR is trained 15 epochs and fine-tuned another 5 epochs. The number of quantization bins is set to 1000. We use DarkNet-53 [39] as the visual encoder. More details are provided in the appendix.

4.4 Comparisons with State-of-the-Arts

In this section, we compare the proposed SeqTR with the state-of-the-art methods on five benchmark datasets, *i.e.*, RefCOCO, RefCOCO+, RefCOCOg, ReferItGame, and Flickr30K Entities. Tab. 1 and Tab. 3 show the performance on

Table 1. Comparison with the state-of-the-arts on the REC task. Visual encoders of models with † is trained without excluding val/test images of the three datasets. RN101 refers to ResNet101 [12] and DN53 denotes DarkNet53 [39].

Models	Visual Encoder	RefCOCO			RefCOCO+			RefCOCOg			Time
		val	testA	testB	val	testA	testB	val-g	val-u	test-u	(ms)
Two-stage											
CMN [14]	VGG16	-	71.03	65.77	-	54.32	47.76	57.47	-	-	-
VC [57]	VGG16	-	73.33	67.44	-	58.40	53.18	62.30	-	-	-
ParalAttn [60]	VGG16	-	75.31	65.52	-	61.34	50.86	58.03	-	-	-
MAttNet [54]	RN101	76.40	80.43	69.28	64.93	70.26	56.00	-	66.58	67.27	320
CM-Att-Erase [29]	RN101	78.35	83.14	71.32	68.09	73.65	58.03	-	67.99	68.67	-
DGA [47]	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28	341
RvG-Tree [13]	RN101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51	-
NMTree [27]	RN101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44	-
One-stage											
RealGIN [58]	DN53	77.25	78.70	72.10	62.78	67.17	54.21	-	62.75	62.33	35
FAOA† [50]	DN53	71.15	74.88	66.32	56.86	61.89	49.46	-	59.44	58.90	39
RCCF [25]	DLA34	-	81.06	71.85	-	70.35	56.32	-	-	65.73	**25**
MCN [32]	DN53	80.08	82.29	74.98	67.16	72.86	57.31	-	66.46	66.01	56
ReSC$_L$† [49]	DN53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20	36
Iter-Shrinking [44]	RN101	-	74.27	68.10	-	71.05	58.25	-	-	70.05	-
LBYL† [16]	DN53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-	30
TransVG [6]	RN101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73	62
TRAR† [59]	DN53	-	81.40	78.60	-	69.10	56.10	-	68.90	68.30	-
SeqTR (ours)	DN53	81.23	85.00	76.08	68.82	75.37	58.78	-	71.35	71.58	50
SeqTR† (ours)	DN53	**83.72**	**86.51**	**81.24**	**71.45**	**76.26**	**64.88**	**71.50**	**74.86**	**74.21**	50

REC and RES tasks. Tab. 4 reports the result of SeqTR pre-trained on the large corpus of data. The performance on ReferItGame and Flickr30K Entities datasets are given in Tab. 2.

The performance of SeqTR on REC and phrase localization tasks is illustrated in Tab. 1 and Tab. 2. From Tab. 1, our model performs better than two-stage models, especially MAttNet [54] while being 6 times faster. We also surpass one-stage models that exploit prior and expert knowledge, with +2-7% absolute improvement over LBYL [16] and ReSC [49]. Despite we predict discrete coordinate tokens in an auto-regressive manner, the inference speed[1] of SeqTR is only 50ms, which is real-time and comparable with one-stage models. For transformer-based models, SeqTR surpasses TransVG [6] and TRAR [59] with up to 6.27% absolute performance improvement. Our SeqTR achieves new state-of-the-art performance with a simple architecture and loss function on the RefCOCO [55], RefCOCO+ [55], and RefCOCOg [34,36] datasets. On the ReferItGame and Flickr30K Entities datasets which mostly contain short noun phrases, the performance boosts to 69.66 and 81.23 with a large margin over previous one-stage methods [50,42,25,49] and is comparable with current state-of-the-art methods [6,24].

[1]Tested on GTX 1080 Ti GPU, batch size is 1.

Table 2. Comparison with the state-of-the-art models on the test set of Flickr30K Entities [37] and ReferItGame [21] datasets.

Models	Visual Encoder	ReferItGame test	Flickr30k test	Time (ms)
Two-stage				
MAttNet [54]	RN101	29.04	-	320
SimilarityNet [46]	RN101	34.54	60.89	184
DDPN [56]	RN101	63.00	73.30	-
One-stage				
FAOA [50]	DN53	60.67	68.71	**23**
ZSGNet [42]	RN50	58.63	63.39	-
RCCF [25]	DLA34	63.79	-	25
ReSC$_L$ [49]	DN53	64.60	69.28	36
TransVG [6]	RN101	70.73	79.10	62
RefTR [24]	RN101	**71.42**	78.66	40
SeqTR (ours)	DN53	69.66	**81.23**	50

SeqTR can be seamlessly extended to RES without any network architecture modifications since we reformulate the task as a point prediction problem. As shown in Tab. 3, we outperform various models with sophisticated cross-modal alignment and reasoning mechanisms [19,31,10,32,51,17,28]. SeqTR is on par with current state-of-the-art VLT [8] which selectively aggregates responses from the diversified queries, whereas we directly produce the corresponding segmentation mask and establish one-to-one correspondence. When initialized with the pre-trained parameters using the large corpus of data, the performance boosts up to 10.78% absolute improvement, proving that a simple yet universal approach for visual grounding is indeed feasible.

From Tab. 4, when pre-trained on the large corpus of text-image pairs, SeqTR is more data-efficient than the current state-of-the-art [20]. Our transformer architecture only contains 7.9M parameters which is twice as few as MDETR [20], while the performance is superior especially on the RefCOCOg dataset with up to 2.48% improvement.

4.5 Ablation Studies

To give a comprehensive understanding of SeqTR, we discuss ablative studies on the validation set of the RefCOCO [55], RefCOCO+ [55], and RefCOCOg [36] datasets in this section.

Construction of language feature. Language feature in Sec. 3.3 can be constructed by either *max*/*mean* pooling of word features or directly using the final hidden state of bi-GRU. As shown in the upper part of Tab. 5, *max* pooling performs best, and is the default construction throughout this paper.

Token weight. If previously predicted points are inaccurate, model can not recover from the wrong predictions since the inference is sequential. Hence we increase a few former token weights to penalize more on the first several predicted

Table 3. Comparison with the state-of-the-arts on the RES task. Model with $*$ is pre-trained on the large corpus of data.

Models	Visual Encoder	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val-g	val-u	test-u
MAttNet [54]	RN101	56.51	62.37	51.70	46.67	52.39	40.08	-	47.64	48.61
CMSA [51]	RN101	58.32	60.61	55.09	43.76	47.60	37.89	39.98	-	-
STEP [2]	RN101	60.04	63.46	57.97	48.19	52.33	40.41	46.40	-	-
BRINet [15]	RN101	60.98	62.99	59.21	48.17	52.32	42.11	48.04	-	-
CMPC [17]	RN101	61.36	64.53	59.64	49.56	53.44	43.23	49.05	-	-
LSCM [18]	RN101	61.47	64.99	59.55	49.34	53.12	43.50	48.05	-	-
CMPC+ [28]	RN101	62.47	65.08	60.82	50.25	54.04	43.47	49.89	-	-
MCN [32]	DN53	62.44	64.20	59.71	50.62	54.99	44.69	-	49.22	49.40
EFN [10]	WRN101	62.76	65.69	59.67	51.50	55.24	43.01	**51.93**	-	-
BUSNet [48]	RN101	63.27	66.41	61.39	51.76	56.87	44.13	50.56	-	-
CGAN [31]	DN53	64.86	68.04	62.07	51.03	55.51	44.06	46.54	51.01	51.69
LTS [19]	DN53	65.43	67.76	63.08	54.21	58.32	48.02	-	54.40	54.25
VLT [8]	DN56	65.65	68.29	62.73	55.50	59.20	49.36	49.76	52.99	56.65
SeqTR (ours)	DN53	67.26	69.79	64.12	54.14	58.93	48.19	-	55.67	55.64
SeqTR* (ours)	DN53	**71.70**	**73.31**	**69.82**	**63.04**	**66.73**	**58.97**	-	**64.69**	**65.74**

Table 4. Comparison with pre-trained models on RefCOCO [55], RefCOCO+ [55], and RefCOCOg [36] datasets. We only count the parameters of transformer architecture.

Models	Visual Encoder	Params (M)	Pre-train images	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val-u	test-u
ViLBERT [30]	RN101	-	3.3M	-	-	-	72.34	78.52	62.61	-	-
VL-BERT$_L$ [43]	RN101	-	3.3M	-	-	-	72.59	78.57	62.30	-	-
UNITER$_L$ [4]	RN101	-	4.6M	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA$_L$ [11]	RN101	-	4.6M	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL$_L$ [53]	RN101	-	4.3M	-	-	-	75.95	82.07	66.88	-	-
MDETR [20]	RN101	17.36	200K	86.75	89.58	81.41	**79.52**	84.09	70.62	81.64	80.89
RefTR [24]	RN101	17.86	**100K**	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
SeqTR (ours)	DN53	**7.90**	174K	**87.00**	**90.15**	**83.59**	78.69	**84.51**	**71.87**	**82.69**	**83.37**

discrete coordinate tokens. As shown in the lower part of Tab. 5, increasing the weight of first token is better than increasing the latter tokens, and setting the 1st token weight to 1.5 and subsequent tokens to 1 gives the best performance. We set $w_i = 1, \forall i$ for RES task.

Sampling scheme. We verify the upper bound as the mIoU of the assembled mask from the sampled points and original ground-truth. From Fig. 4 (a), we can see that the mIoU approaches nearly 100 when the number of sampled points increases, *i.e.*, 95.57 for uniform sampling, and 91.58 for center-based sampling. Therefore, though the upper bound is limited theoretically, in practice, the research effort might be better spent on improving the real-world performance. In terms of sampling strategies, from Fig. 4 (a) and Fig. 4 (c-e), uniform sampling is consistently better than center-based sampling in terms of both the upper bound and the performance, which preserves more details of the mask illustrated in Fig. 3. The number of sampled points controls the trade-off between

Table 5. Ablation experiments on the construction of language feature and token weight. The first token is the [TASK] token, while subsequent tokens are discrete coordinate tokens, *i.e.*, (x_1, y_1, x_2, y_2).

Language feature	Token weight 1st	2nd	3rd	4th	5th	RefCOCO val	RefCOCO+ val	RefCOCOg val-u
mean pooling	1	1	1	1	1	79.73	67.12	68.97
max pooling						**80.07**	**68.31**	**69.95**
final hidden state						79.85	67.46	69.93
max pooling	1	1	1	1	1	80.07	68.31	69.95
	1.5	1	1	1	1	80.10	**68.63**	**70.05**
	2	1	1	1	1	**80.19**	68.33	70.01
	3	1	1	1	1	80.08	67.81	69.45
	1	2	2	1	1	79.70	67.22	69.51
	2	2	2	1	1	80.16	67.83	69.45

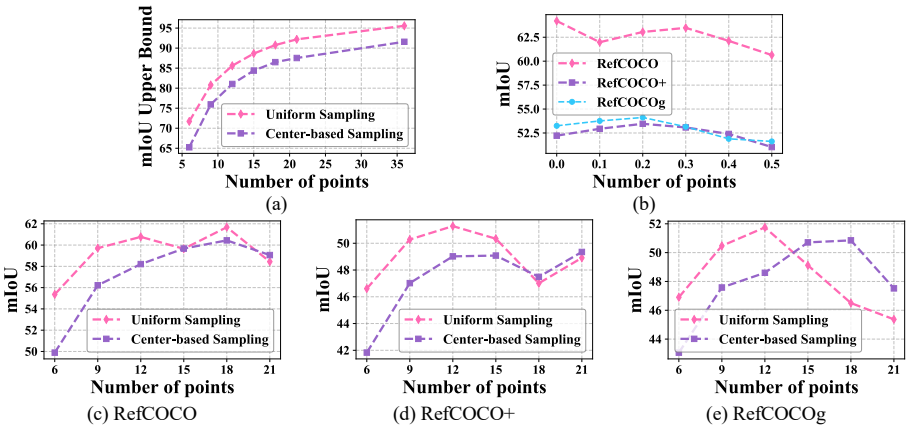

Fig. 4. Ablative experiments on RES task. (a) The upper bound is averaged over validation sets (the fluctuation is within 0.2). (b) Shuffling percentage refers to the fraction of shuffled sequences within a batch, uniform sampling strategy is used. (c-e) depict the impact of sampling strategies and the number of sampled points.

the inference speed and performance, from Fig. 4 (c-e), we can see that 18 and 12 points are the best for RefCOCO and RefCOCO+/RefCOCOg datasets.

Shuffling percentage. We train SeqTR 60 epochs instead of 30 as we empirically found that point shuffling takes a longer time to converge, since the ground-truth is different for each coordinate token at each forward pass. Fig. 4 (b) shows that no shuffle and 0.2 are best for RefCOCO and RefCOCO+/RefCOCOg. As the number of shuffled sequences increases, the performance drops slightly, and we observe that SeqTR is under-fitting since the mIoU during training is lower than the one without shuffling.

Multi-task training. Previous multi-task visual grounding approaches require REC to help RES locate the referent. In contrast, *SeqTR is capable to locate*

Table 6. Ablation study of multi-task training. IE is the inconsistency error [32] to measure the prediction conflict between REC and RES, ↓ denotes the lower is better.

Dataset	Multi-task training	REC	RES				
		Prec@0.5	Prec@0.5	Prec@0.7	Prec@0.9	mIoU	IE↓
RefCOCO	✗	**80.38**	**78.03**	**63.35**	**9.75**	**64.20**	13.93
	✓	79.65	77.24	60.29	7.23	62.93	**5.86**
RefCOCO+	✗	67.98	65.11	48.27	5.19	52.22	22.22
	✓	**68.79**	**66.67**	**51.02**	**5.46**	**53.65**	**4.85**
RefCOCOg	✗	69.63	**65.20**	**46.23**	**5.31**	**53.25**	22.65
	✓	**70.29**	**65.20**	46.05	5.15	**53.25**	**8.25**

the referent at pixel level without the aid from REC. We train SeqTR 60 epochs and test whether multi-task supervision can bring further improvement. For the input sequence construction of multi-task grounding, please see the supplementary material. From Tab. 6, we can see that multi-task supervision even slightly degenerates the performance compared to the single-task variant. Though the inconsistency error significantly decreases, the location ability of RES measured by Prec@0.5, 0.7, and 0.9 stays the same, suggesting that the sampled points are independent between the sequence of the bounding box and binary mask.

4.6 Qualitative Results

We visualize the cross attention map averaged over decoder layers and attention heads in Fig. 5. At each prediction step, SeqTR generates a coordinate token given previous output tokens. Under this setting, a clear pattern emerges, *i.e.*, attends to the left side of the referent when predicting x_1, the top side of the referent when predicting y_1, and so on. This axial attention is sensitive to the boundary of the referent, thus can more precisely ground the referred object. The predicted masks are visualized in Fig. 6. SeqTR can well comprehends attributive words and absolute or relative spatial relations, and the predicted mask aligns with the irregular outlines of the referred object such as "*left cow*". More qualitative results are given in the appendix.

5 Conclusions

In this paper we reformulate visual grounding tasks as a point prediction problem and present an innovative and general network termed SeqTR. Based on the standard transformer encoder-decoder architecture and *cross-entropy* loss, SeqTR unifies different visual grounding tasks under the same point prediction paradigm without any modifications. Experimental results demonstrate that SeqTR can well ground language query onto the corresponding region, suggesting that a simple yet universal approach for visual grounding is indeed feasible.

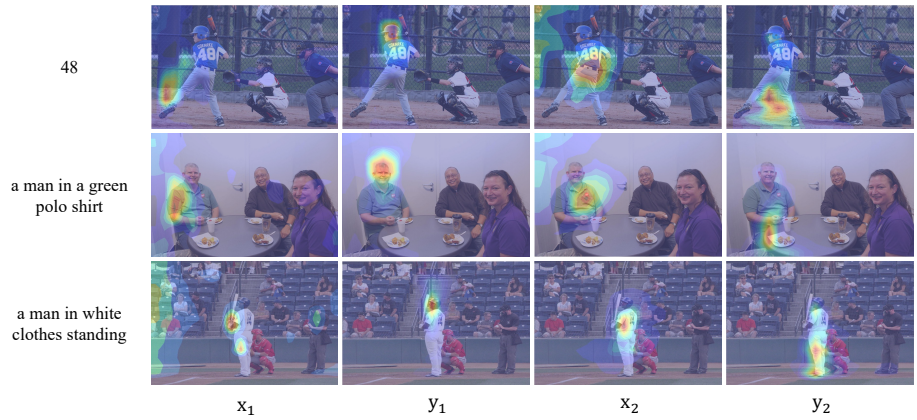

Fig. 5. Visualization of normalized cross attention map in transformer decoder. From left to right column, we generate (x_1, y_1, x_2, y_2) in sequential order.

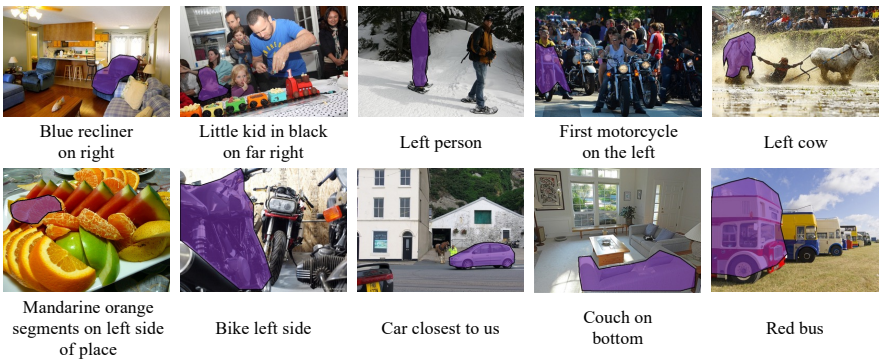

Fig. 6. Example mask predictions by SeqTR on the validation set of RefCOCO dataset, best viewed in color.

Acknowledgements. This work was supported by the National Science Fund for Distinguished Young Scholars (No. 62025603), the National Natural Science Foundation of China (No. U21B2037, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, and No. 62002305), Guangdong Basic and Applied Basic Research Foundation (No. 2019B1515120049), and the Natural Science Foundation of Fujian Province of China (No. 2021J01002).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 213–229 (2020)
2. Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7454–7463 (2019)
3. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. In: International Conference on Learning Representations (ICLR) (2022)
4. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 104–120 (2020)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
6. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1769–1779 (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16321–16330 (2021)
9. Escalante, H.J., Hernández, C.A., Gonzalez, J.A., López-López, A., Montes, M., Morales, E.F., Sucar, L.E., Villasenor, L., Grubinger, M.: The segmented and annotated iapr tc-12 benchmark. Computer Vision and Image Understanding (CVIU) **114**(4), 419–428 (2010)
10. Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15506–15515 (2021)
11. Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6616–6628 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
13. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
14. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1115–1124 (2017)
15. Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H.: Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4424–4433 (2020)

16. Huang, B., Lian, D., Luo, W., Gao, S.: Look before you leap: Learning landmark features for one-stage visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16888–16897 (2021)
17. Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10488–10497 (2020)
18. Hui, T., Liu, S., Huang, S., Li, G., Yu, S., Zhang, F., Han, J.: Linguistic structure guided context modeling for referring image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 59–75 (2020)
19. Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., Tan, T.: Locate then segment: A strong pipeline for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9858–9867 (2021)
20. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1780–1790 (2021)
21. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 787–798 (2014)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017)
24. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)
25. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10880–10889 (2020)
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017)
27. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4673–4682 (2019)
28. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2021)
29. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1950–1959 (2019)
30. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)

31. Luo, G., Zhou, Y., Ji, R., Sun, X., Su, J., Lin, C.W., Tian, Q.: Cascade grouped attention network for referring expression segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia (MM). pp. 1274–1282 (2020)
32. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10034–10043 (2020)
33. Luo, G., Zhou, Y., Sun, X., Ding, X., Wu, Y., Huang, F., Gao, Y., Ji, R.: Towards language-guided visual recognition via dynamic convolutions. arXiv preprint arXiv:2110.08797 (2021)
34. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11–20 (2016)
35. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
36. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 792–807. Springer (2016)
37. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Internatioanl Journal of Computer Vision (IJCV) **123**(1), 74–93 (2017)
38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
39. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS) **28** (2015)
41. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 658–666 (2019)
42. Sadhu, A., Chen, K., Nevatia, R.: Zero-shot grounding of objects from natural language queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4694–4703 (2019)
43. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
44. Sun, M., Xiao, J., Lim, E.G.: Iterative shrinking for referring expression grounding using deep reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14060–14069 (2021)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)
46. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **41**(2), 394–407 (2018)

47. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4644–4653 (2019)
48. Yang, S., Xia, M., Li, G., Zhou, H.Y., Yu, Y.: Bottom-up shift and reasoning for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11266–11275 (2021)
49. Yang, Z., Chen, T., Wang, L., Luo, J.: Improving one-stage visual grounding by recursive sub-query construction. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 387–404 (2020)
50. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4683–4693 (2019)
51. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10502–10511 (2019)
52. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics (TACL) **2**, 67–78 (2014)
53. Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H.: Ernie-vil: Knowledge enhanced vision-language representations through scene graph. arXiv preprint arXiv:2006.16934 (2020)
54. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
55. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 69–85 (2016)
56. Yu, Z., Yu, J., Xiang, C., Zhao, Z., Tian, Q., Tao, D.: Rethinking diversified and discriminative proposal generation for visual grounding. arXiv preprint arXiv:1805.03508 (2018)
57. Zhang, H., Niu, Y., Chang, S.F.: Grounding referring expressions in images by variational context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4158–4166 (2018)
58. Zhou, Y., Ji, R., Luo, G., Sun, X., Su, J., Ding, X., Lin, C.W., Tian, Q.: A real-time global inference network for one-stage referring expression comprehension. IEEE Transactions on Neural Networks and Learning Systems (TNNLS) (2021)
59. Zhou, Y., Ren, T., Zhu, C., Sun, X., Liu, J., Ding, X., Xu, M., Ji, R.: Trar: Routing the attention spans in transformer for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2074–2084 (2021)
60. Zhuang, B., Wu, Q., Shen, C., Reid, I., Van Den Hengel, A.: Parallel attention: A unified framework for visual object discovery through dialogs and queries. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4252–4261 (2018)