Improving Visual Grounding with Visual-Linguistic Verification and Iterative Reasoning

Li Yang[1,2*], Yan Xu[3*], Chunfeng Yuan[1†], Wei Liu[1], Bing Li[1], and Weiming Hu[1,2,4]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]The Chinese University of Hong Kong
[4]CAS Center for Excellence in Brain Science and Intelligence Technology

{li.yang,cfyuan,bli,wmhu}@nlpr.ia.ac.cn, yanxu@link.cuhk.edu.hk, liuwei@ia.ac.cn

Abstract

Visual grounding is a task to locate the target indicated by a natural language expression. Existing methods extend the generic object detection framework to this problem. They base the visual grounding on the features from pre-generated proposals or anchors, and fuse these features with the text embeddings to locate the target mentioned by the text. However, modeling the visual features from these predefined locations may fail to fully exploit the visual context and attribute information in the text query, which limits their performance. In this paper, we propose a transformer-based framework for accurate visual grounding by establishing text-conditioned discriminative features and performing multi-stage cross-modal reasoning. Specifically, we develop a visual-linguistic verification module to focus the visual features on regions relevant to the textual descriptions while suppressing the unrelated areas. A language-guided feature encoder is also devised to aggregate the visual contexts of the target object to improve the object's distinctiveness. To retrieve the target from the encoded visual features, we further propose a multi-stage cross-modal decoder to iteratively speculate on the correlations between the image and text for accurate target localization. Extensive experiments on five widely used datasets validate the efficacy of our proposed components and demonstrate state-of-the-art performance.

1. Introduction

Visual grounding aims to localize the referred object or region in an image by a natural language expression. This task has received increasing attention because of its great potential in bridging the gap between linguistic expressions and visual perception. The evolution of this technique is

---

∗ denotes equal contribution.
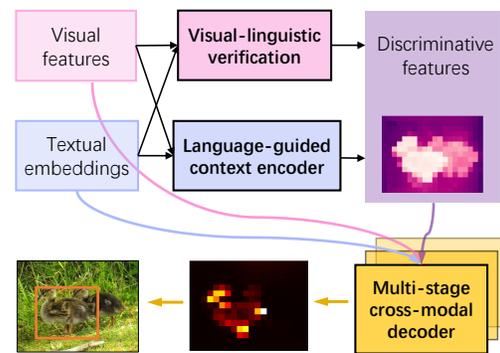† denotes the corresponding author.

Figure 1. Our proposed framework for visual grounding. With the features from the two modalities as input, the visual-linguistic verification module and language-guided context encoder establish discriminative features for the referred object. Then, the multi-stage cross-modal decoder iteratively mulls over all the visual and linguistic features to identify and localize the object.

also of great importance to other multi-modal reasoning tasks. In visual grounding, the referred object is generally specified by one or more pieces of information in the language expression. The information may include object categories, appearance attributes, and visual relation contexts, *etc*. Thus, to avoid ambiguity in reasoning, it is crucial to fully exploit the textual information and model discriminative visual features for visual grounding.

Existing methods, no matter the two-stage or one-stage ones, treat visual grounding as a ranking problem on the detected candidate regions. The two-stage methods [23, 24, 33, 45] generally first detect a set of object proposals and then match them with the language query to retrieve the top-ranked proposal. The one-stage approaches [3, 18, 43] directly fuse the text embeddings with image features to generate dense detections, from which to choose the one with the highest confidence score. As these methods are based on generic object detectors, their inference procedures rely on

the predictions from all possible candidate regions, which makes the performance limited by the quality of the pre-predicted proposals or the configuration of predefined anchor boxes. Moreover, they represent the candidate objects with the region features (corresponding to the predicted proposals) or the point features (of the dense anchor boxes) to match or fuse with the text embeddings. Such feature representations may be less flexible for capturing detailed visual concepts or contexts mentioned in linguistic descriptions. This inflexibility may increase the difficulties in recognizing the target object. Although some methods exploit modular attention [44], graph and tree [20, 34, 40, 41] to better model the relations between vision and language, their processing pipelines are complicated and the performance is still limited by the object proposal inputs.

Recently, the boom of the transformer in natural language processing [6, 32] and computer vision [1, 7] has shown its powerful modeling capability in both the language and vision fields. Motivated by that, TransVG [5] proposes a transformer-based framework for visual grounding. Taking the visual and linguistic feature tokens as inputs, they stack a set of transformer encoder layers to perform cross-modal fusion, and directly predict the target locations. Despite its effectiveness, the shared transformer encoder layers are in charge of multiple tasks, including encoding the visual-linguistic features, identifying the object instances, and acquiring the final locations, which may increases the learning difficulty and could only achieve compromised results. It is also less straightforward to retrieve the visual features of the target object with their feature fusion mechanism. Thus, we propose to establish a more dedicated framework for accurate visual grounding.

In this work, we propose a transformer-based visual grounding framework that directly retrieves the target object's feature representation for localization. To this end, as shown in Fig. 1, we first establish the discriminative feature representations by visual-linguistic verification and context aggregation, and then identify the referred object by multi-stage reasoning. Specifically, the visual-linguistic verification module compares the visual features with the semantic concepts from textual embeddings to focus on the regions relevant to the language expression. In parallel, the language-guided context encoder gathers the context features to make the visual features of the target object more distinguishable. Based on these enhanced features, we propose a multi-stage cross-modal decoder that iteratively compares and mulls over the visual and linguistic information. This enable us to progressively acquire a better representation of the referred object and thereby determine its location more accurately.

To summarize, our contributions are three-fold: (1) To establish the discriminative features for visual grounding, we propose a visual-linguistic verification module to focus the encoded features on the regions related to the language expression. A language-guided context encoder is

further devised to aggregate important visual contexts for better object identification. (2) To retrieve a more accurate feature representation of the referred object, we propose a multi-stage cross-modal decoder, which iteratively queries and mulls over visual and linguistic information to reduce the ambiguity during inference. (3) We benchmark our method on RefCOCO [45], RefCOCO+ [45], RefCOCOg [23], ReferItGame [16], and Flickr30k Entities [26]. Our method exhibits significant performance improvements over the previous state-of-the-art methods. Extensive experiments and ablation studies validate the efficacy of our proposed components. Our code is public at https://github.com/yangli18/VLTVG.

2. Related Work

2.1. Visual Grounding

Existing methods generally extend the object detection framework [10, 27, 28, 39, 48] to address the visual grounding task. Two-stage methods [12, 13, 20, 33, 34, 40, 44, 47, 50] leverage the off-the-shelf detectors to generate a set of proposals from the image in the first stage, and then match them with the language expression to select the top-ranked proposal. However, these methods heavily rely on the performance of the pre-trained detector or proposal generator. If the referred object is not detected in the first stage, the ranking and selection process in the second stage will also fail to output the correct detection.

Recently, one-stage approaches [3, 18, 42, 43] predict the location of referred object without generating the candidate proposals in advance. For instance, FAOA [43] encodes the language expression into a textual embedding and fuse it into the YOLOv3 detector [27]. The model generates dense object detections with confidence scores, and select the top-ranked one as the prediction for the referred object. In order to solve the shortcomings in modeling long language queries, ReSC [42] proposes a iterative sub-query construction framework to reduce the referring ambiguity. Despite their efficiency, the one-stage methods generally utilize the point feature as the object representation, which may be less flexible to associate with the detailed descriptions in the language expression.

2.2. Visual Transformer

While convolutional neural networks (CNNs) achieved promising results in various vision tasks [11, 28, 35–38], the success of transformers in both vision and language fields has attracted great attention of the research community, and the transformer has emerged as a viable alternative to CNNs in many vision tasks, such as image classification [7], object detection [1, 49], *etc*. For example, DETR [1] proposes an end-to-end detection framework based on the transformer framework. This work utilizes the attention mechanism of transformer to formulate object detection as a set prediction problem, which obtains outstanding performance. De-
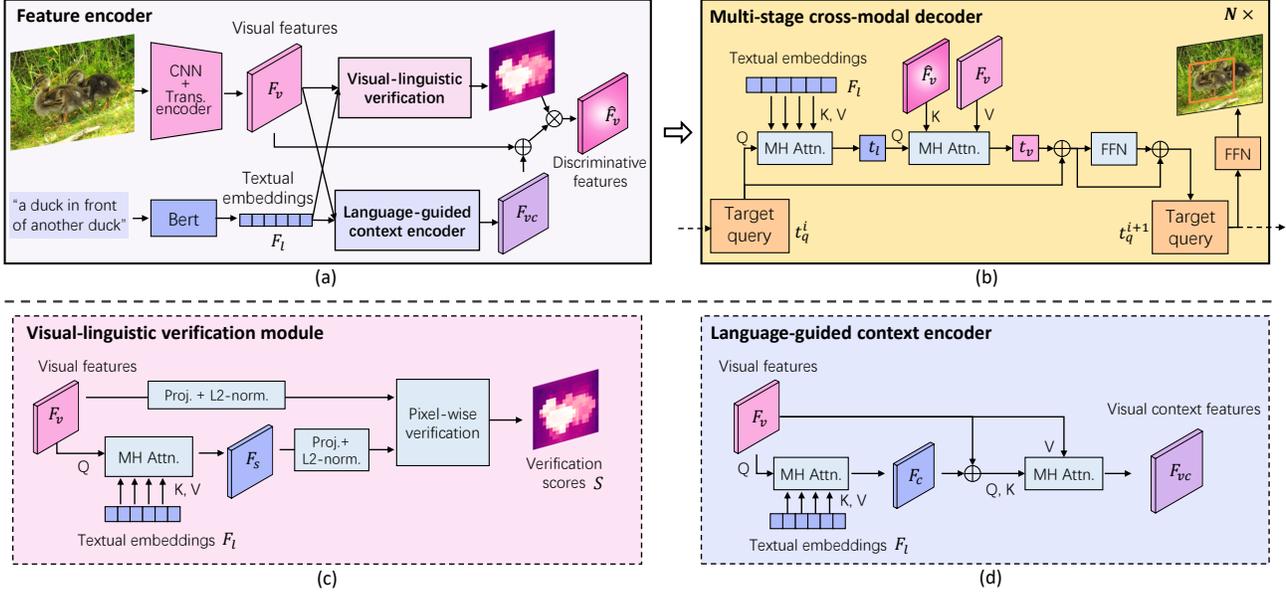
Figure 2. The overall network architecture of our proposed method. Given the input image and language expression, the feature encoder (a) first extracts the visual features and textual embeddings, respectively. Then, these features are processed by the visual-linguistic verification module (c) and language-guided context encoder (d) to produce more discriminative features for the referred object. Finally, the multi-stage cross-modal decoder (b) leverages all the visual and textual features to iteratively infer the target location.

formable DETR [49] introduces deformable convolution to reduce the computational cost of DETR. ViT [7] shows that a transformer-based backbone can achieve excellent performance on image classification. All these applications in vision tasks demonstrate the powerful and general modeling capabilities of transformer. Motivated by this, we also propose to utilize transformer to develop a more flexible and effective visual grounding framework.

There are also works exploiting transformer modules or attention mechanisms to solve different types of vision and language tasks [4, 17, 22, 31]. For example, SCAN [17] addresses image-text matching by modeling correlations in the proposal-level. STVGBert [31] correlates the text embedding with the video frame features for video grounding. In contrast, we focus on visual grounding based on images, and we model the visual-linguistic correlations pixel-wisely to focus feature encoding on the semantically related regions for grounding.

3. Method

In this section, we present our proposed framework for visual grounding. We first introduce the overall network architecture. Then, we elaborate on our proposed visual-linguistic verification module, language-guided context encoder and multi-stage cross-modal decoder. Finally, we detail the loss function for training.

3.1. The Overall Network

Unlike the previous ranking-based visual grounding methods, we directly retrieve the object feature represen-

tation for object localization. As shown in Fig. 2, given an image and a language expression, we first input them into two separate branches for feature extraction. For the image, we utilize a convolutional neural network (*e.g.*, ResNet-50 [11]) with a stack of transformer encoder layers to generate a 2D visual feature map $F_v \in \mathbb{R}^{C \times H \times W}$. For the input language expression, we encode it as a sequence of textual embeddings $F_l \in \mathbb{R}^{C \times L}$ with BERT [6] in the other branch. Then, based on the features of both modalities, we apply the visual-linguistic verification module and language-guided context encoder to encode discriminative features $\hat{F}_v \in \mathbb{R}^{C \times H \times W}$ for the referred object. The visual-linguistic verification module refines the visual features to focus on the regions related to the referring expression, while the language-guided context encoder gathers informative visual contexts to facilitate the target-object identification. Finally, a multi-stage cross-modal decoder is applied to iteratively mull over the encoded visual and textual features to more exactly retrieve the object representation for target localization.

3.2. Visual-Linguistic Verification Module

The input image is encoded first by the convolutional network and then the follow-up transformer encoder layers to be a visual feature map F_v. This feature map contains the features of object instances in the image, but has no prior knowledge about the referred object by the language expression. Retrieving the representation of the referred object without any prior could be distracted by other objects or regions, causing less accurate localization results. To ad-

dress this issue, we propose the visual-linguistic verification module to compute fine-grained correlations between the visual and textual features and focus features on regions related to the semantics in the textual description.

As shown in Fig. 2-(c), the visual-linguistic verification module is based on the multi-head attention [32]. Here, the visual feature map $F_v \in \mathbb{R}^{C \times H \times W}$ serves as the query and the textual embeddings $F_l \in \mathbb{R}^{C \times L}$ acts as the key and value. With the multi-head attention, the related semantic features will be gathered (from the textual embeddings) for each visual feature vector of the visual feature map F_v. The gathered semantic features are organized as a semantic map $F_s \in \mathbb{R}^{C \times H \times W}$ that is spatially aligned with the visual feature map F_v. Thereafter, we project both the feature maps F_v and F_s to the same semantic space (via linear projection and L2 normalization) obtaining F_v' and F_s', and compute their semantic correlation as the verification score for each spatial location (x, y) as:

$$S(x, y) = \alpha \cdot \exp\left(-\frac{\left(1 - F_v'(x, y)^{\mathrm{T}} F_s'(x, y)\right)^2}{2\sigma^2}\right), \quad (1)$$

where α and σ are learnable parameters. The verification scores model the semantic relevance of each visual feature to the linguistic expression. Thereby, we are able to establish more salient feature map $\hat{F}_v$ for the referred object by modulating the visual features with the verification scores pixel-wisely:

$$\hat{F}_v = F_v \cdot S . \quad (2)$$

Such modulated visual features $\hat{F}_v$ naturally suppress regions that are irrelevant to the referring expression, making it easier to recognize and locate the referred object in the later phases.

3.3. Language-guided Context Encoder

In addition to establishing the semantic correlations between visual and linguistic representations, modeling the visual contexts (*e.g.*, the interaction relations and relative positions) is also crucial for distinguishing the referred target object from other parts. To this end, we propose a language-guided context encoder to gather context features under the guidance of the textual descriptions.

As illustrated in Fig. 2-(d), based on the multi-head attention module (left-most one), we input the visual feature map F_v as the query to collected the relevant semantic information from the textual embeddings, producing the feature map F_c as the corresponding linguistic representation for the visual features. Based on F_c, we employ another multi-head attention module to perform language-guided self-attention for the visual features to encode the involved visual contexts. Specifically, we take the sum of F_v and F_c as the query and key of the multi-head attention module, to compute the feature correlations in both visual and linguistic representations. For each attention head in the multi-head attention module, we formulate the attention value

from position i to another position j as:

$$\begin{cases} Q = W_Q^{\mathrm{T}}(F_v + F_c) \\ K = W_K^{\mathrm{T}}(F_v + F_c) \\ \mathrm{attn}_{i,j} = \mathrm{softmax}(\dfrac{Q(i)^{\mathrm{T}}(K(j) + W_K^{\mathrm{T}} R(i-j))}{\sqrt{d_k}}) \end{cases}, \quad (3)$$

where W_Q and W_K are the linear projection weights for the query and key, d_k is the projection channel dimension, and $R(\cdot)$ is the sinusoidal positional encodings [32] of relative positions. This language-guided self-attention enables the model to learn to gather important context features for the referred object based on the given textual descriptions. The multi-head attention module outputs the map of context feature representations as F_{vc}. To establish more discriminative features for the target object, we extend the Eq. (2) to fuse both the context features F_{vc} and visual-linguistic verification scores S (from Eq. (1)) with the visual features F_v as:

$$\hat{F}_v = (F_v + F_{vc}) \cdot S . \quad (4)$$

The produced discriminative feature representations are then utilized in the final multi-stage decoder for target identification and localization.

3.4. Multi-stage Cross-modal Decoder

We leverage the established visual feature maps and the textual embeddings for final target localization. To reduce the ambiguity in reasoning, we propose a multi-stage cross-modal decoder that iteratively mulls over the visual and linguistic information to distinguish the target object from other parts and retrieve related features for object localization.

We illustrate the architecture of the decoder in Fig. 2-(b). The decoder consists of N stages, each of which is constituted by the same network architecture (with unshared weights) for iterative cross-modal reasoning. In the first stage, we employ a learnable target query $t_q^1 \in \mathbb{R}^{C \times 1}$ as the initial representation of the target object. This target query is fed into the decoder to extract visual features based on the linguistic expression, and iteratively updates its feature representation into t_q^i $(1 \leq i \leq N)$ in the following stages. Fig. 2-(b) shows the feature updating process of each decoder stage. Specifically, in the i-th stage, the target query t_q^i is input (as the query) to the first multi-head attention module to collect informative semantic descriptions about the referred object from the textual embeddings, producing the linguistic representation $t_l \in \mathbb{R}^{C \times 1}$. Then, in the second multi-head attention module, we use this linguistic representation t_l as the query to compute its correlation with the previously established discriminative features $\hat{F}_v$ (from the feature encoder, Fig. 2-(a)), and then gather the features of interest from the visual feature map F_v. In this manner, the gathered visual feature $t_v \in \mathbb{R}^{C \times 1}$ for the referred object is produced based on the collected semantic descriptions in

t_l. Thereafter, we use this gathered feature t_v to update the target query t_q^i as:

$$\begin{cases} t_q' = \mathrm{LN}(t_q^i + t_v) \\ t_q^{i+1} = \mathrm{LN}(t_q' + \mathrm{FFN}(t_q')) \end{cases} \quad (5)$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization, and $\mathrm{FFN}(\cdot)$ is a feed-forward network composed of two linear projection layers with ReLU activations. The updated target query t_q^{i+1} is then input to the next decoder stage for iterative cross-modal reasoning and feature representation updating.

Based on this multi-stage structure, the target query t_q^i of each stage is able to focus on different descriptions in the referring expression, and thus more complete and reliable features of the target object can be gathered. The target query t_q^i is progressively refined with the gathered features to form a more accurate representation of the target object. Finally, we append a three-layer MLPs with ReLU activation functions to the target query output in each stage to predict the bounding box of the referred object. The predicted bounding boxes from all stages are equally supervised to facilitate the multi-stage decoder training.

3.5. Training Loss

Our training procedure is more concise and straightforward than the previous ranking-based methods [42, 44]. As our network directly regresses the final bounding box, we avoid the positive/negative sample assignment and can directly compute the regression losses with the ground-truth box. Let $\{\hat{b}^i\}_{i=1}^N$ denote the predicted bounding boxes from stages 1 to N, and b denote the ground-truth box. We sum up the losses between the predicted boxes and the ground-truth box over all the stages as:

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_{\mathrm{giou}} \mathcal{L}_{\mathrm{giou}}(b, \hat{b}^i) + \lambda_{\mathrm{L1}} \mathcal{L}_{\mathrm{L1}}(b, \hat{b}^i) , \quad (6)$$

where $\mathcal{L}_{\mathrm{giou}}(\cdot, \cdot)$ and $\mathcal{L}_{\mathrm{L1}}(\cdot, \cdot)$ are the GIoU loss [29] and L1 loss respectively. The λ_{giou} and λ_{L1} are the hyperparameters to balance the two losses.

4. Experiments

4.1. Datasets

RefCOCO/ RefCOCO+/ RefCOCOg. RefCOCO [45], RefCOCO+ [45], and RefCOCOg [23] are three commonly used benchmarks for visual grounding, with images collected from the MS COCO dataset [19]. 1) RefCOCO [45] has 19,994 images with 142,210 referring expressions for 50,000 referred objects. It is split into train, validataion, testA, and testB sets with 120,624, 10,834, 5657, and 5095 expressions, respectively. 2) RefCOCO+ [45] provides 19,992 images with 49,856 referred objects and 141,564 referring expressions. It is also split into the train, validation, testA, and testB sets that have 120,191, 10,758,

5,726, and 4,889 referring expressions, respectively. 3) RefCOCOg [23] has 25,799 images with 95,010 referring expressions for 49,822 object instances. The expressions in RefCOCOg are generally longer than those in the other two datasets. There are two splitting conventions for RefCOCOg, which are RefCOCOg-google [23] (val-g) and RefCOCOg-umd [24] (val-u, test-u). We conduct experiments following the two conventions separately for comprehensive comparisons.

ReferItGame. ReferItGame [16] contains 20,000 images collected from the SAIAPR-12 dataset [8]. We follow the previous works [14,42] to split the dataset into three subsets, *i.e.*, a train set with 54,127 referring expressions, a validation set with 5,842 referring expressions, and a test set with 60,103 referring expressions.

Flickr30k Entities. Flickr30k Entities [26] has 31,783 images with 427k referred entities. We follow the same split as in the previous works [5, 26, 42] for training, validation, and testing.

4.2. Implementation Details

We set the input image size to 640×640 and the maximum length of the language expression to 40. During inference, we resize the input image to make the longer edge equal to 640 and pad the shorter edge to 640. We append the [CLS] and [SEP] tokens to the beginning and end of the input language expression respectively, before processing the expression with BERT [6].

During training, we use the AdamW optimizer [21] to train our proposed model with a batch size of 64. We set the initial learning rate of our network to 10^{-4}, except for the feature extraction branches (*i.e.* the CNN + transformer encoder layers, and BERT), which have an initial learning rate of 10^{-5}. We use ResNet-50 or ResNet-101 as our CNN backbone followed by 6 transformer encoder layers in the visual feature extraction branch, which are initialized with the corresponding weights of DETR model [1]. The textual embedding extraction branch is initialized with BERT [6]. We use Xavier init [9] to randomly initialize the parameters of the other components in our network. We train our model for 30 epochs in all ablation studies, where the learning rate decays by a factor of 10 after 20 epochs. For comparison with state-of-the-art methods, we extend the training epochs to 90, and decay the learning rate by 10 after 60 epochs. To stabilize the training, we also freeze the weights of the visual and textual feature extraction branches in the first 10 epochs. For the learnable parameters in Eq. (1), we set $\alpha = 1.0$ and $\sigma = 0.5$ as the initial values. For the loss function in Eq. (6), we set $\lambda_{\mathrm{giou}} = 2$ and $\lambda_{\mathrm{L1}} = 5$. We follow the previous works [5, 18, 42, 43] to perform data augmentation during training.

We follow the evaluation metric of the previous works [5, 42]. Given an image and a language expression, the predicted bounding box is considered correct if its IoU with the ground-truth bounding box is larger than 0.5.

Table 1. Comparison of our method with other state-of-the-art methods on RefCOCO [45], RefCOCO+ [45], and RefCOCOg [23].

Models	Venue	Backbone	RefCOCO			RefCOCO+			RefCOCOg		
			val	testA	TestB	val	testA	testB	val-g	val-u	test-u
Two-stage:											
CMN [13]	CVPR'17	VGG16	-	71.03	65.77	-	54.32	47.76	57.47	-	-
VC [47]	*CVPR'18*	VGG16	-	73.33	67.44	-	58.40	5.318	62.30	-	-
ParalAttn [50]	*CVPR'18*	VGG16	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MAttNet [44]	*CVPR'18*	ResNet-101	76.65	81.14	69.99	65.33	71.62	56.02	-	66.58	67.27
LGRANs [34]	*CVPR'19*	VGG16	-	76.60	66.40	-	64.00	53.40	61.78	-	-
DGA [40]	*ICCV'19*	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28
RvG-Tree [12]	*TPAMI'19*	ResNet-101	75.06	78.61	69.85	63.51	67.45	56.66	-	66.95	66.51
NMTree [20]	*ICCV'19*	ResNet-101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44
Ref-NMS [2]	*AAAI'21*	ResNet-101	80.70	84.00	76.04	68.25	73.68	59.42	-	70.55	70.62
One-stage:											
SSG [3]	*arXiv'18*	DarkNet-53	-	76.51	67.50	-	62.14	49.27	47.47	58.80	-
FAOA [43]	*ICCV'19*	DarkNet-53	72.54	74.35	68.50	56.81	60.23	49.60	56.12	61.33	60.36
RCCF [18]	*CVPR'20*	DLA-34	-	81.06	71.85	-	70.35	56.32	-	-	65.73
ReSC-Large [42]	*ECCV'20*	DarkNet-53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20
LBYL-Net [15]	*CVPR'21*	DarkNet-53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-
Transformer-based:											
TransVG [5]	*ICCV'21*	ResNet-50	80.32	82.67	78.12	63.50	68.15	55.63	66.56	67.66	67.44
TransVG [5]	*ICCV'21*	ResNet-101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73
VLTVG (ours)		ResNet-50	**84.53**	**87.69**	**79.22**	**73.60**	**78.37**	**64.53**	**72.53**	**74.90**	**73.88**
VLTVG (ours)		ResNet-101	**84.77**	**87.24**	**80.49**	**74.19**	**78.93**	**65.17**	**72.98**	**76.04**	**74.18**

4.3. Comparisons with State-of-the-art Methods

In Table 1, we report our performance in comparison with other state-of-the-art methods on RefCOCO [45], RefCOCO+ [45], and RefCOCOg [23] datasets. Our method outperforms other methods in all splits of the three datasets. Notably, we achieve absolute improvements of up to 4.45%, 5.94% and 5.49% respectively on RefCOCO, RefCOCO+, and RefCOCOg, compared with the cutting-edge two-stage method Ref-NMS [2]. The recent two-stage methods [2, 44] rely on a pre-trained object detector, *e.g.*, Mask R-CNN [10] for object proposal generation and feature extraction. Since their object detector is not involved in visual grounding training, the visual features from the detector may not be very compatible with the visual grounding task. Moreover, the quality of the pre-generated proposals could also become the performance bottleneck of these two-stage paradigms. In contrast, our method encodes the target's feature representation with the guidance from the language expression, which is more flexible and adaptive.

Our method also shows consistent improvements over the current one-stage methods [15, 42]. Notably, we achieve accuracies of 84.77% and 80.49% on the val and testB splits of RefCOCO, which bring absolute improvements of 5.10 and 6.34 percentage points over the previous best performance of the one-stage framework [15]. The previous one-stage methods use the fused visual-textual feature at each spatial location to represent a candidate object, which may not adequately capture the important visual attributes or contexts in the language expression. Instead, our method models the representation of the target object by iteratively querying the linguistic and visual features, allowing us to identify the referred object in a finer-grained way.

Comparing with the the most recent work TransVG [5], our method also achieves appreciable improvements. As shown in Table 1, our accurices exceed TransVG by $2.14\% \sim 4.52\%$ on RefCOCO, $8.23\% \sim 9.37\%$ on RefCOCO+, and $5.96\% \sim 7.37\%$ on RefCOCOg.

Table 2 also reports the performance of our method on the test sets of the ReferItGame [16] and Flickr30k Entities [26]. Our method surpasses the previous one-stage and two-stage methods by an appreciable margin. Compared with TransVG [5], we achieve relatively better performance with all the different backbones. We notice that our improvements over TransVG on Flickr30k Entities are less than the improvements on the other datasets. This could be because most of the referring expressions in Flickr30k Entities are short noun phrases, which may be not suitable for exhibiting the advantages of our method. Nevertheless, the experimental results demonstrate the generality and competitiveness of our method in different scenarios.

4.4. Ablation Study

In this section, we conduct the ablation studies on the RefCOCOg (val-g) [23] dataset, containing longer language expressions, which poses more challenges for comprehension and reasoning.

4.4.1 The Component Modules

In Table 3, we conduct a thorough ablation study on the proposed components to verify their effectiveness. The first row of Table 3 shows our baseline that performs visual grounding with only a single-stage decoder, which achieves an accuracy of 63.64%. Based on this baseline, we further

Table 2. Comparison with the state-of-the-art methods on the test sets of ReferItGame [16] and Flickr30k Entities [26].

Models	Backbone	ReferItGame test	Flickr30K test
Two-stage:			
CMN [13]	VGG16	28.33	-
VC [47]	VGG16	31.13	-
MAttNet [44]	ResNet-101	29.04	-
Similarity Net [33]	ResNet-101	34.54	60.89
CITE [25]	ResNet-101	35.07	61.33
DDPN [46]	ResNet-101	63.00	73.30
One-stage:			
SSG [3]	DarkNet-53	54.24	-
ZSGNet [30]	ResNet-50	58.63	63.39
FAOA [43]	DarkNet-53	60.67	68.71
RCCF [18]	DLA-34	63.79	-
ReSC-Large [42]	DarkNet-53	64.60	69.28
LBYL-Net [15]	DarkNet-53	67.47	-
Transformer-based:			
TransVG [5]	ResNet-50	69.76	78.47
TransVG [5]	ResNet-101	70.73	79.10
VLTVG (ours)	ResNet-50	**71.60**	**79.18**
VLTVG (ours)	ResNet-101	**71.98**	**79.84**

Table 3. The ablation studies of the proposed components in our network. We evaluate the accuracy of visual grounding, and report the model size and the computational complexity.

Multi-stage Decooder	Context Encoder	V-L Verification	#params	GFLOPS	Acc (%)
			143.37M	41.10	63.64
✓			151.26M	41.39	66.02
✓	✓		151.79M	41.67	68.44
✓	✓	✓	152.18M	41.79	**71.62**

add more cross-modal decoder stages to perform iterative reasoning for target localization. It can be found that the accuracy is notably improved by 2.38 points, as shown in the second row of Table 3. In the third row of Table 3, we further introduce the language-guided context encoder to gather visual contexts for grounding, and find the accuracy is again boosted by 2.42 points. The last row of Table 3 shows the performance of our full model after applying the visual-linguistic verification module, which further improves the accuracy from 68.44% to 71.62% (+3.18 percentage points) and achieves the best performance among these ablation variants.

We also report the number of model parameters and the computational complexity of different variants in Table 3. Our proposed components only introduce a total of 8.81M extra parameters and 0.69 GFLOPS to the baseline, which only increase the model size and computational complexity by 6.14% and 1.68%, respectively.

4.4.2 The Decoder Stages

In Table 4, we evaluate our system when employing different numbers of decoder stages for the object location decod-

Table 4. Comparison of different decoder stages used to perform cross-modal reasoning for visual grounding.

The decoder stages (N)	#params	GFLOPS	Acc (%)
$N = 1$	143.37M	41.10	63.64
$N = 2$	144.95M	41.15	65.05
$N = 4$	148.10M	41.27	65.70
$N = 6$	151.26M	41.39	**66.02**
$N = 8$	154.42M	41.51	65.97

Table 5. Comparison of our method with the transformer-based approach for visual-linguistic feature learning.

The V-L feature learning	#params	GFLOPS	Acc (%)
None	151.26M	41.39	66.02
Trans. encoder layers ($\times 1$)	152.69M	42.08	69.37
Trans. encoder layers ($\times 2$)	154.00M	42.77	69.22
Trans. encoder layers ($\times 3$)	155.32M	43.46	69.15
Trans. encoder layers ($\times 4$)	156.63M	44.15	69.55
Ours (V-L verification + context)	152.18M	41.79	**71.62**

ing. As shown in Table 4, the accuracy steadily increases (from 65.22%) as more decoder stages are employed until it reaches the saturation point of 67.57% near $N = 6$. This reflects the importance of multi-stage reasoning for visual grounding, The multi-stage decoder queries the linguistic information and gathers the visual features with multiple rounds, allowing the referred object to be identified and localized more accurately. As the accuracy improves little when $N \geq 6$, we employ 6 decoder stages in our network by default. Besides, our multi-stage decoder only increases the model parameters by 5.5% and the computational cost by 0.71%, which is quite efficient.

4.4.3 Visual-Linguistic Feature Learning

In our network, we utilize the visual-linguistic verification module and the language-guided context encoder for feature learning of both modalities. To further verify the necessity and effectiveness, we compare our method with the common transformer-based design for visual-linguistic feature learning in Table 5. Specifically, we replace our verification module and context encoder with a stack of common transformer encoder layers. We concatenate the visual and textual features and feed them to the stacked transformer encoder layers to perform cross-modal feature fusion. As shown in Table 5, using a single transformer encoder layer for feature learning improves the accuracy by 3.35 points over the baseline. However, when stacking more transformer encoder layers, we observe no significant improvements in accuracy, but a increase in the number of parameters and computational cost. In contrast, our method improves the baseline by 5.60 percentage points with fewer parameters and computation resources than a single transformer encoder layer. Our method also outperforms the
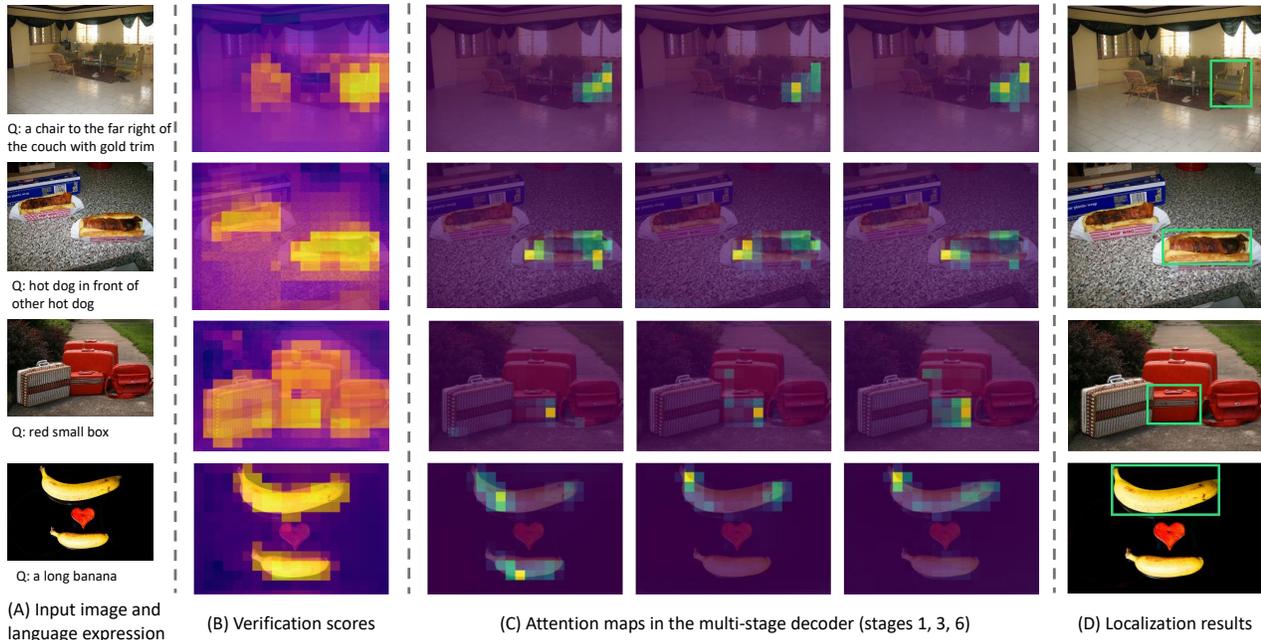
Figure 3. Visualization of the verification score maps, the multi-stage decoder's attention maps, and the final localization results for various input images and language expressions.

Figure 4. Visualization of the attention map of the language-guided context encoder.

top variant with transformer encoders by 2.07 percentage points, which demonstrates the efficacy of our design in visual-linguistic features encoding.

4.5. Visualization

In Fig. 3, we visualize the generated verification scores, the multi-stage decoder's attention maps, and the final localization results for different inputs. It can be observed that, the verification scores are generally higher on the objects or regions related to the descriptions in the language expressions. With Eq. (2), our modulated visual features are able to focus on these regions and reduce the distractions. Then, the multi-stage decoder mulls over these regions to localize the referred object. For example, in the 4-th row of Fig. 3, given the language query "a long banana", the attention of the first stage is focused on both two bananas. After multi-stage reasoning, the attention is successfully drawn to the longer banana mentioned by the text, for localizing the target. The discriminative feature representation and multi-stage reasoning enable our method to effectively locate the referred objects for various image and expression inputs.

In Fig. 4, given a language expression, we visualize the

language-guided context encoder's attention map for a point on the referred girl. Guided by the text, the context encoder focuses on the context around the related horse. The gathered visual contexts are helpful for identifying the girl in the image.

5. Conclusions and Limitations

We have presented a transformer-based visual grounding framework, which establishes discriminative features and performs iterative cross-modal reasoning for accurate target localization. Our visual-linguistic verification module focuses the visual feature encoding on the regions related to the textual descriptions, while the language-guided context encoder gathers informative visual contexts to improve the distinctiveness of the target. Moreover, the multi-stage cross-modal decoder iteratively mulls over the visual and linguistic information for localization. Extensive experiments on the public datasets exhibit the state-of-the-art performance of our method.

One limitation of this work is that our model is only trained on the visual grounding datasets with limited sizes of corpus. Our trained model may not generalize well to more open language expressions. In the future, we plan to extend our method to larger scale grounding datasets for better generalization ability.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 5

[2] Long Chen, Wenbo Ma, Jun Xiao, Hanwang Zhang, and Shih-Fu Chang. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1036–1044, 2021. 6

[3] Xinpeng Chen, Lin Ma, Jingyuan Chen, Zequn Jie, Wei Liu, and Jiebo Luo. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*, 2018. 1, 2, 6, 7

[4] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 3

[5] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 2, 5, 6, 7

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 5

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2, 3

[8] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villasenor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010. 5

[9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 5

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[12] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 6

[13] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124, 2017. 2, 6, 7

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4555–4564, 2016. 5

[15] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. 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*, pages 16888–16897, 2021. 6, 7

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 5, 6, 7

[17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 3

[18] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 1, 2, 5, 6, 7

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[20] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 2, 6

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[23] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 2, 5, 6

[24] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 1, 5

[25] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018. 7

[26] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2, 5, 6, 7

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 2

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 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*, pages 658–666, 2019. 5

[30] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019. 7

[31] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 3

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 4

[33] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018. 1, 2, 7

[34] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 2, 6

[35] Yan Xu, Zhaoyang Huang, Kwan-Yee Lin, Xinge Zhu, Jianping Shi, Hujun Bao, Guofeng Zhang, and Hongsheng Li. Selfvoxelo: Self-supervised lidar odometry with voxel-based deep neural networks. In *Conference on Robot Learning*, pages 115–125. PMLR, 2021. 2

[36] Yan Xu, Junyi Lin, Jianping Shi, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Robust self-supervised lidar odometry via representative structure discovery and 3d inherent error modeling. *IEEE Robotics and Automation Letters*, 2022. 2

[37] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization, 2022. 2

[38] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2811–2820, 2019. 2

[39] Li Yang, Yan Xu, Shaoru Wang, Chunfeng Yuan, Ziqi Zhang, Bing Li, and Weiming Hu. Pdnet: Towards better one-stage object detection with prediction decoupling. *arXiv preprint arXiv:2104.13876*, 2021. 2

[40] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 2, 6

[41] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961, 2020. 2

[42] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 2, 5, 6, 7

[43] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 2, 5, 6, 7

[44] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2, 5, 6, 7

[45] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1, 2, 5, 6

[46] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 7

[47] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018. 2, 6, 7

[48] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020. 2

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 2, 3

[50] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018. 2, 6