

Revisiting Counterfactual Problems in Referring Expression Comprehension

Zhihan Yu and Ruifan Li*

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
 {yzh0, rfl1}@bupt.edu.cn

Abstract

Traditional referring expression comprehension (REC) aims to locate the target referent in an image guided by a text query. Several previous methods have studied on the Counterfactual problem in REC (C-REC) where the objects for a given query cannot be found in the image. However, these methods focus on the overall image-text or specific attribute mismatch only. In this paper, we address the C-REC problem from a deep perspective of fine-grained attributes. To this aim, we first propose a fine-grained counterfactual sample generation method to construct C-REC datasets. Specifically, we leverage pre-trained language model such as BERT to modify the attribute words in the queries, obtaining the corresponding counterfactual samples. Furthermore, we propose a C-REC framework. We first adopt three encoders to extract image, text and attribute features. Then, our dual-branch attentive fusion module fuses these cross-modal features with two branches by an attention mechanism. At last, two prediction heads generate a bounding box and a counterfactual label, respectively. In addition, we incorporate contrastive learning with the generated counterfactual samples as negatives to enhance the counterfactual perception. Extensive experiments show that our framework achieves promising performance on both public REC datasets RefCOCO+/+g and our constructed C-REC datasets C-RefCOCO+/+g. The code and data are available at <https://github.com/Glacier0012/CREC>.

1. Introduction

Given an image and a text query, referring expression comprehension (REC) [37, 42] aims to locate the visual target referent guided by the query. An example is shown in Figure 1 a). REC connects image regions with natural language, which contributes to downstream vision-language tasks such as image captioning [3, 16, 49] and visual question answering [11, 41, 45]. In addition, REC task has various practical applications, including interactive photo edit-

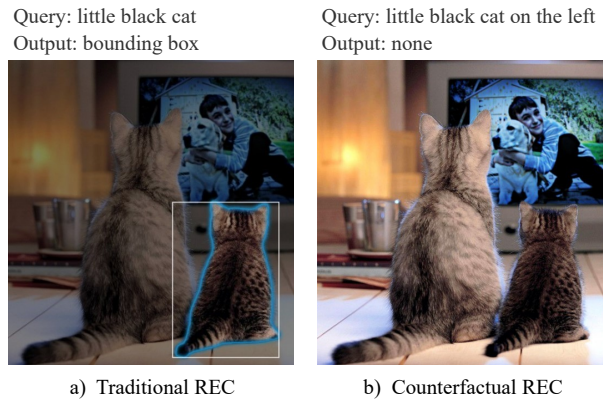


Figure 1. Comparison of traditional REC and counterfactual REC. a) In traditional REC, the target referent guided by text query always exists in the given image, which is marked by a bounding box. b) In counterfactual REC, the target referent cannot be found.

ing [44] and robot navigation [1, 36].

Generally, REC models can be grouped into two categories, two-stage models [5, 25, 26, 50] and one-stage models [7, 27, 46, 48, 51]. The former group extracts region proposals from off-the-shelf object detectors [12, 35] and ranks them by semantic similarity to output the top-score region. In contrast, the latter group has an end-to-end structure with faster inference speed and competitive performance. However, most of these two groups of methods suppose that the target referent can always be found in the given image. They overlook the counterfactual text queries that could appear in real-world scenarios. For example, when a home service robot is asked to grasp “a fork” but there is only “a knife” in its sight, the bounding box output by most REC models will cause wrong movements. In fact, a more reasonable reaction of the robot is to provide a “no target” feedback. We refer to this counterfactual problem as Counterfactual Referring Expression Comprehension (C-REC).

Several methods have recently studied on C-REC. One group considers C-REC as a matching task based on extracted entity and specific attribute [6, 10]. The other adopts classification frameworks that detect the overall counterfac-

*Corresponding author.

tual polarity of image-text pairs [22, 24]. However, these methods ignore the image-text mismatch on various fine-grained attributes. For example, in Figure 1 b), the little black cat is on the *right*. Thus, the query “little black cat on the *left*” is a fine-grained counterfactual query. This query considers the attribute of location. Motivated by above observations, we propose to revisit the C-REC problem from a deep perspective of fine-grained attributes.

Naturally, two questions arise for addressing C-REC task. **Firstly**, how to generate fine-grained counterfactual samples effectively? Fine-grained counterfactual samples generation requires modification on words of various attributes (see Table 1). A naive idea of manually modifying keywords in texts is labor-intensive and costly to collect large-scale datasets. Besides, a simple automatic method is to design replacement rules based on a pre-defined vocabulary, such as mutual replacement of the size attribute words “big” and “small”. However, it is difficult to cover all the attributes words. In short, effectively generating C-REC datasets should be addressed. **Secondly**, how to learn joint cross-modal features for simultaneously performing localization and counterfactual detection? The former task inclines towards global features from text queries. In contrast, the latter task prefers local textual features of attributes. The seemingly contradictory requirements for the learned features make fine-grained C-REC challenging. In addition, easily available counterfactual samples are unique for C-REC task. Therefore, using these counterfactual samples wisely would benefit for learning the joint features.

To tackle these problems, **firstly** we propose a counterfactual sample generation (CSG) method to synthesize counterfactual text queries in a labor-free way. We use a dependency parsing tool [30] to extract words of pre-defined attributes for a query from existing REC datasets. Then we mask these words to predict new ones by pre-trained language models. To obtain counterfactual predictions, we adopt a re-ranking scheme to exclude synonyms. Using our CSG method, we build three fine-grained counterfactual datasets C-RefCOCO+/g. **Secondly**, we propose a C-REC framework with resilience to counterfactual queries. The framework learns cross-modal joint features using a dual-branch attentive fusion (DAF) module. Furthermore, we incorporate contrastive learning wisely using the generated counterfactual samples to enhance counterfactual perception. At last, we conduct extensive experiments, obtaining promising performance of around 90% counterfactual classification accuracy on C-RefCOCO+/g and improvements on box accuracy on RefCOCO+/g. Ablation studies prove the effective designs in our C-REC framework.

Our major contributions are summarized as follows.

1) We revisit the counterfactual problem in REC from a deep perspective of fine-grained attributes. We propose a counterfactual sample generation method to build C-REC

Table 1. Fine-grained attributes and their corresponding words in the referring expression, “*small yellow wooden boat on the river in the center of the image*”.

| ID | Attribute | Word |
|----|----------------------------|---------------|
| A1 | head noun | <i>boat</i> |
| A2 | color | <i>yellow</i> |
| A3 | size | <i>small</i> |
| A4 | absolute location relation | <i>center</i> |
| A5 | relative location relation | <i>on</i> |
| A6 | relative location object | <i>river</i> |
| A7 | generic attribute | <i>wooden</i> |

datasets in a labor-free way.

2) We propose a C-REC framework to detect the counterfactual polarity and simultaneously to locate the target referents. We incorporate dual-branch attentive fusion and contrastive learning to enhance counterfactual perception.

3) We conduct extensive experiments to show that our framework achieves promising performance on both REC datasets and our constructed C-REC datasets.

2. Methodology

We formulate the problem of counterfactual referring expression comprehension (C-REC) as a multi-task framework composed of binary classification and coordinate regression. Given an image I and a text query T , the goal of C-REC task is to predict a counterfactual label $C = \{0, 1\}$ and simultaneously to locate the target referent with a bounding box B . Note that $C = 1$ indicates the query and the image are matched pairs, and $C = 0$ indicates the query is counterfactual for the image. The box B is a vector of (x, y, w, h) , where (x, y) represents the center point of the box, and w and h represent the width and height of the box. Next, we will describe our counterfactual sample generation method and then propose our C-REC model.

2.1. Counterfactual Sample Generation (CSG)

Our C-REC samples are based on fine-grained attributes in referring expressions. Inspired by ReferItGame [18], we define seven types of attributes, including *head noun*, *color*, *size*, *absolute location relation*, *relative location relation*, *relative location object*, and *generic attribute*. Specifically, *head noun* is the center noun of the referring expression which indicates the category of target referent. *Absolute location relation* refers to the location of the target in the image. *Relative location relation* refers to the location of the target referent relative to another object which is noted as *relative location object*. *Generic attribute* includes the general appearance features that are less frequently observed such as material, shape and state. These attributes cover the most common appearance and spatial information of target

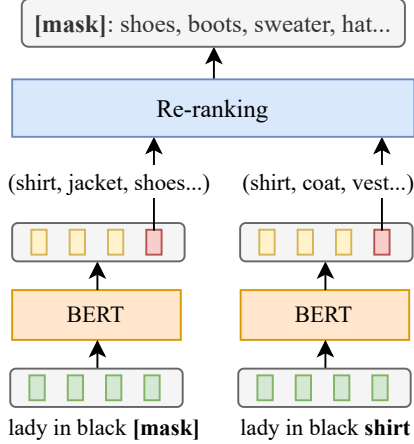


Figure 2. Counterfactual Sample Generation (CSG) method. An attribute word is extracted by dependency parsing and masked. Then, BERT is used to predict candidate words for [MASK] token and we re-rank them to generate counterfactual queries.

referents. Table 1 shows the fine-grained attributes and corresponding words in a referring expression.

Moreover, to obtain fine-grained C-REC samples, we generate negative texts [13] based on existing REC datasets. We leverage a language model to modify the attribute words to counterfactual ones, while preserving the context of the text query. The overview of our proposed counterfactual sample generation (CSG) method is shown in Figure 2. Specifically, the CSG method consists of the following four steps. **1)** We extract all the attribute words in given text query T from existing REC datasets using a dependency parsing tool [30]. Note that the seven attributes cannot always be found in a query, and in most cases there only exist two or three attributes. For text queries of multiple attributes, we generate one counterfactual sample for each attribute word. After that, the attribute word a is replaced by a [MASK] token. **2)** We leverage the pre-trained language model such as BERT [8] to predict N candidate words for the [MASK] token. These words are most likely to appear according to the context \mathcal{C}_{ct} . Every candidate word obtains a probability $P(w_i|\mathcal{C}_{ct})$, where $i = 1, 2, \dots, N$. **3)** The initial query T is fed into BERT. We obtain the probability of candidate words on the position of a given the context \mathcal{C}_{ct} and the attribute word a , which is denoted as $P(w_i|\mathcal{C}_{ct}, a)$. Intuitively, among these candidate words, the attribute word a has the maximum probability, followed by some synonyms. **4)** To select the counterfactual words from the predictions, we adopt a re-ranking scheme. We define a correlation score s_i between the candidate word w_i and the attribute word a as follows,

$$s_i = \frac{P(w_i|\mathcal{C}_{ct})}{P(w_i|\mathcal{C}_{ct}, a)}, \quad i = 1, 2, \dots, N. \quad (1)$$

We re-rank these words by the correlation score s_i and replace a with the re-ranked candidates. To promote the robustness of the counterfactual query, we randomly choose one of the top N' re-ranked candidate words as the negative sample of query T . The selected new word appears in context \mathcal{C}_{ct} with a high probability but appears in context \mathcal{C}_{ct} and a with a low probability. This indicates a low correlation with the attribute word a . Thus, we discard the synonyms of the initial attribute word. In other words, the generated text queries are semantically counterfactual in comparison with the initial query.

2.2. Counterfactual REC Model

The overview of the proposed model is shown in Figure 3. Our C-REC model has a one-stage structure, consisting of three encoders, a dual-branch attentive fusion (DAF) module, a regression head and a counterfactual detection head. In addition, contrastive learning and overall loss are described. Details of our model are elaborated as follows.

Encoders. Our model adopts three encoders to extract features from images, text queries and attribute words, respectively. We take CSPDarkNet [40] as image encoder to capture visual information of diverse semantic levels. Feature maps from the last K layers are output as visual features. These features are denoted as $F_v^i \in R^{k_i \times k_i \times d_i}$, $i = 1, \dots, K$, where d_i and k_i are the dimension and map size of the i -th layer. For the text encoding, we take LSTM [15] with GLOVE embeddings [34] as text encoder. The query T is padded and pooled into a global vector. The textual feature is denoted as $F_t \in R^{d_t}$, where d_t represents the textual feature dimension. At last, the attribute encoder shares the parameters with the text encoder. Thus, the attribute word is encoded into the attribute feature $F_a \in R^{d_t}$.

Dual-branch Attentive Fusion (DAF). To effectively utilize the linguistic encodings of different granularity, we adopt a two-branch structure for cross-modal fusion. Inspired by RealGIN [51], we apply an attention mechanism for each branch. In one branch, visual feature F_v and textual feature F_t are projected to an identical dimension. These features are fused by dot-product to obtain a scalar product f_{vt} as follows,

$$f_{vt}^i = \sigma(W_v F_v^i)^T \cdot \sigma(W_t F_t), \quad (2)$$

where W_v and W_t are projection matrices and $\sigma(\cdot)$ is Sigmoid function. Then we calculate the attention map a_c and sum it into F_v to obtain the attention feature f_{att} , i.e.,

$$a_c^i = \frac{\exp(f_{vt}^i)}{\sum_j \exp(f_{vt}^j)}, \quad f_{att} = \sum_i a_c^i F_v^i. \quad (3)$$

Furthermore, to diffuse the attention feature f_{att} into the fusion feature with the identical dimension of F_v , we cal-

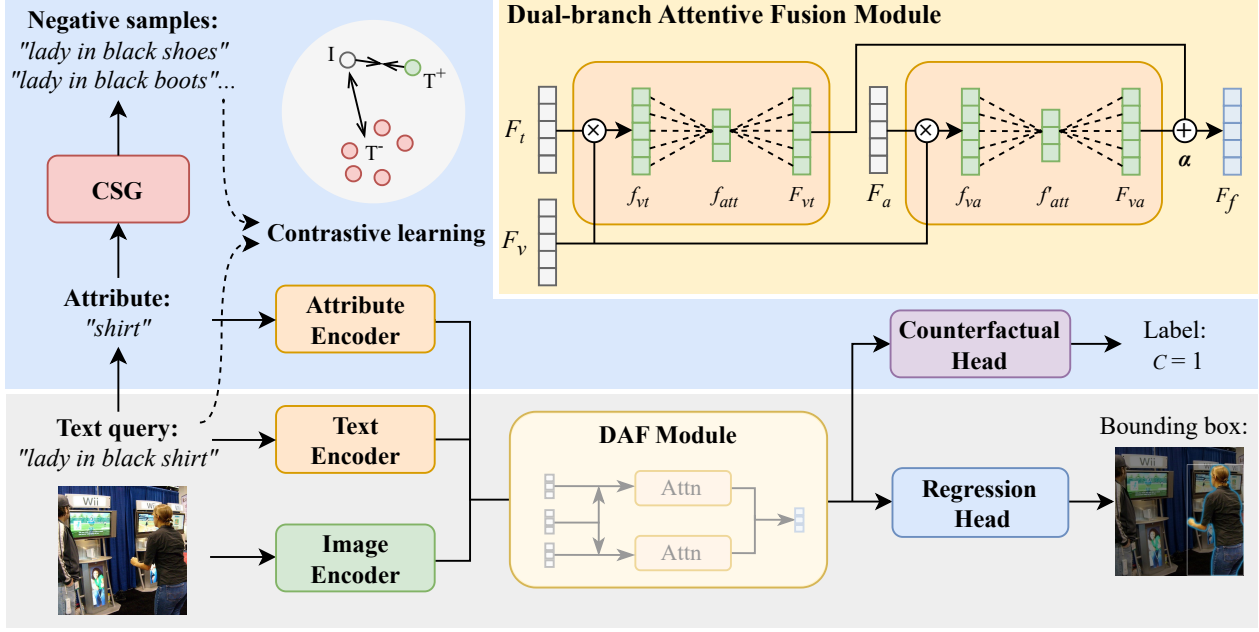


Figure 3. Overview of our C-REC model. We first adopt three encoders to extract the image, text and attribute features, respectively. Then we fuse the three features in the dual-branch attentive fusion (DAF) module to obtain a fusion feature. After that, a regression head is used to predict the bounding box and a counterfactual head is used to predict the counterfactual label. In addition, we incorporate a contrastive loss using the negative samples generated by CSG method for enhancing cross-modal fusion.

culate a diffusion attention map a_d . This attention map is calculated in the same way as that of the attention map a_c . After that, we obtain the image-text fusion feature F_{vt} with a residual connection as follows,

$$F_{vt}^i = F_v^i + a_d^i f_{att}. \quad (4)$$

Similarly, the other branch fuses the visual feature F_v and attribute feature F_a as the above branch does. Thus, we obtain the image-attribute feature F_{va} , which guides the visual maps to attend to the crucial attribute information in the query. Lastly, the fusion feature F_f used for prediction is obtained by averaging features from the previous two branches. This process is formulated as follows,

$$F_f = \alpha F_{vt} + (1 - \alpha) F_{va}, \quad (5)$$

where the parameter α is used to balance the dual branches. Note that the layer number i is omitted for simplicity.

Prediction Heads. Our model has two independent prediction heads to generate two outputs: a bounding box and a counterfactual label. For localization, the coordinates b and the confidence score c of the bounding box are predicted by convolutional regression layers. The localization loss is given as follows,

$$\mathcal{L}_{loc} = c' \cdot \mathcal{L}_{iou}(b, b') + \mathcal{L}_{ce}(c, c'), \quad (6)$$

where b' and c' are the ground truth of bounding box and

confidence distribution. The term \mathcal{L}_{iou} is IoU loss and the term \mathcal{L}_{ce} is cross-entropy loss.

For counterfactual prediction, we set a binary classifier. The fusion feature F_f is pooled into a fixed-size global vector by an average pooling function, then projected to a 2-D vector through two fully-connected layers with ReLU. We obtain the predicted probability p by Softmax function. The counterfactual classification loss \mathcal{L}_{cf} follows the cross-entropy loss,

$$\mathcal{L}_{cf} = \mathcal{L}_{ce}(p, p'), \quad (7)$$

where p' is the one-hot vector of ground truth label.

Contrastive Learning. To improve the counterfactual perceptual ability of our model, we introduce the contrastive loss. Intuitively, counterfactual text queries could be used as hard negative samples. To this aim, a natural idea is using our proposed CSG method. Our goal is to minimize the distance between an image I and its positive text query T^+ as well as to maximize the distance between the image I and its negative text queries T^- in the latent space. We take InfoNCE loss [14] as contrastive loss, given as follows,

$$\mathcal{L}_{cl} = -\log \frac{\exp(F_f \cdot F_{T^+}/\tau)}{\sum_{t' \in (T^+, T^-)} \exp(F_f \cdot F_{t'}/\tau)}, \quad (8)$$

where τ is the temperature parameter.

Training and Inference. During the training stage, the

Table 2. Statistics on RefCOCO+/g for REC task and our built datasets C-RefCOCO+/g for C-REC task. The number of normal and counterfactual samples in C-RefCOCO+/g is 1:1.

| Dataset | Train | Val | TestA (Test) | TestB |
|------------|-------|-------|--------------|-------|
| RefCOCO | 42404 | 10834 | 5657 | 5095 |
| RefCOCO+ | 42278 | 10758 | 5726 | 4889 |
| RefCOCOg | 42226 | 4896 | 9602 | - |
| C-RefCOCO | 61870 | 15566 | 6994 | 8810 |
| C-RefCOCO+ | 59962 | 15328 | 7846 | 7108 |
| C-RefCOCOg | 30298 | 3676 | 7122 | - |

Table 3. Statistics on fine-grained attributes in C-RefCOCO+/g.

| Dataset | A1 | A2 | A3 | A4 | A5 | A6 | A7 |
|------------|--------|--------|-------|--------|-------|-------|-------|
| C-RefCOCO | 23862 | 5136 | 464 | 16142 | 131 | 131 | 754 |
| C-RefCOCO+ | 28573 | 9864 | 1685 | 2646 | 0 | 0 | 2354 |
| C-RefCOCOg | 11312 | 4114 | 638 | 4024 | 108 | 108 | 244 |
| Percentage | 58.26% | 15.64% | 2.24% | 20.76% | 0.17% | 0.17% | 2.76% |

overall loss is the weighted sum of three terms,

$$\mathcal{L} = \mathcal{L}_{loc} + \gamma_{cf}\mathcal{L}_{cf} + \gamma_{cl}\mathcal{L}_{cl}, \quad (9)$$

where γ_{cf} and γ_{cl} are hyper-parameters.

During the inference stage, our model will output the counterfactual label as well as the bounding box with the highest confidence score.

3. Experiments

3.1. Datasets

To evaluate our C-REC framework, we adopt three REC benchmark datasets and build C-REC datasets. The three REC datasets include RefCOCO, RefCOCO+ and RefCOCOg (briefly denoted as RefCOCO+/g). They are built upon the images from MS-COCO [23]. RefCOCO and RefCOCO+ [32] are created in a two-player game. They both have train, validation and two test splits. TestA contains people instances and TestB contains object ones. The main difference is that RefCOCO includes descriptions of spatial relations while RefCOCO+ forbids them. RefCOCOg [31] contains longer and more complex expressions on appearances and locations compared to two previous datasets. It is split into train/val/test sets.

In addition, we use the proposed CSG method to generate C-REC datasets, named C-RefCOCO+/g for evaluation. In CSG, we set the number of predictions $N = 10$ and the number of re-ranked candidates $N' = 5$. For the dataset construction, we select part of the original data by the query length or extracted attributes to generate negative samples. Then, we add the corresponding positive samples into our datasets to keep the number of normal and counterfactual samples balanced. The size of REC and C-REC datasets are shown in Table 2. The number and proportion



| Image | Text query |
|--|---|
|  | <p>Original query: <u>lady</u> in <u>black</u> on the right of the red umbrella</p> <p>Counterfactual query:</p> <ol style="list-style-type: none"> <u>men</u> in black on the right of the red umbrella lady in <u>green</u> on the right of the red umbrella |
|  | <p>Original query: the <u>pizza</u> on the <u>left</u> more square looking</p> <p>Counterfactual query:</p> <ol style="list-style-type: none"> the <u>window</u> on the left more square looking the pizza on the <u>stove</u> more square looking |

Figure 4. Fine-grained counterfactual samples in C-RefCOCO. The counterfactual text queries are generated by replacing the initial attribute words with semantically opposite or irrelevant words. Head nouns are in red and the other attribute words are in blue.

of different attributes in counterfactual samples are shown in Table 3. In addition, several counterfactual samples in C-RefCOCO are shown in Figure 4.

3.2. Metrics

To evaluate our model, we use two typical metrics and another our proposed metric. The first metric is **Acc-Box** (IoU@0.5). It evaluates the localization performance by measuring the percentage of bounding boxes whose IoU is greater than 0.5. The second metric is **Acc-Cls**. It evaluates the counterfactual detection performance by measuring the percentage of correct predictions on counterfactual labels.

In addition, we design a new metric, i.e., **Acc-Cf** for evaluating the overall performance of a C-REC model. Specifically, we define true positives (TP) as the normal samples whose predicted labels $C = 1$ and simultaneously $IoU > 0.5$, while true negatives (TN) are the counterfactual samples whose predicted labels $C = 0$. Thus, Acc-Cf measures the percentage of correct predictions in N test samples as follows,

$$Acc-Cf = \frac{N_{TP} + N_{TN}}{N}. \quad (10)$$

3.3. Implementation Details

All methods are implemented on two GPUs of NVIDIA RTX A6000. Adam [20] is applied as the optimizer. Batch size is set to 32 and the initial learning rate is 0.0001. The localization result B is enabled only when the counterfactual label is $C = 1$. We set bounding box $B = (0, 0, 0, 0)$ if $C = 0$ for simplicity. The image encoder CSPDarkNet [40] is pre-trained on MS-COCO [23] without the images in validation and test sets. We train our model on RefCOCO+/g for first 40 epochs, and then fine-tune on counterfactual datasets C-RefCOCO+/g for another 20 epochs. The temperature parameter τ in contrastive loss is set to 0.2,

Table 4. Acc-Box (%) of our model and baseline models on RefCOCO+/g. Best results are in **bold** and sub-optimal results are underlined.

| Model | Visual Encoder | Pretrained Images | RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---------------------------------|----------------|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | val | testA | testB | val | testA | testB | val-u | test-u |
| Vision-Language Pretrain | | | | | | | | | | |
| MDETR[17] | ResNet-101 | 200K | 86.75 | 89.58 | 81.41 | 79.52 | 84.09 | 70.62 | 81.64 | 80.89 |
| OFA[43] | ResNet-152 | 20M | 90.05 | 92.93 | 85.26 | 85.80 | 89.87 | 79.22 | 85.89 | 86.55 |
| m-PLUG[21] | ViT-L | 14M | 92.40 | 94.51 | 88.42 | 86.02 | 90.17 | 78.17 | 85.88 | 86.42 |
| One-stage REC | | | | | | | | | | |
| FAOA[46] | DarkNet-53 | - | 72.54 | 74.35 | 68.50 | 56.81 | 60.23 | 49.60 | 61.33 | 60.36 |
| ReSC[47] | DarkNet-53 | - | 77.63 | 80.45 | 72.30 | 63.59 | 68.36 | 56.81 | 67.30 | 67.20 |
| MCN[28] | DarkNet-53 | - | 80.08 | 82.29 | 74.98 | 67.16 | 72.86 | 57.31 | 66.46 | 66.01 |
| RealGIN[51] | DarkNet-53 | - | 77.25 | 78.70 | 72.10 | 62.78 | 67.17 | 54.21 | 62.75 | 62.33 |
| LG-FPN[39] | DarkNet-53 | - | 82.07 | 84.66 | 77.63 | 69.97 | 76.40 | <u>61.32</u> | 71.73 | 71.13 |
| PFOS[38] | DarkNet-53 | - | 79.50 | 81.49 | 77.13 | 65.76 | 69.61 | 60.30 | 69.06 | 68.34 |
| VGTR[9] | ResNet-101 | - | 79.30 | 82.16 | 74.38 | 64.40 | 70.85 | 55.84 | 66.83 | 67.28 |
| TransVG[7] | ResNet-101 | - | 81.02 | 82.72 | 78.35 | 64.82 | 70.70 | 56.94 | 68.67 | 67.73 |
| SimREC[29] | CSPDarkNet-53 | - | <u>82.45</u> | <u>85.91</u> | <u>77.98</u> | <u>70.58</u> | <u>76.75</u> | 61.12 | <u>72.59</u> | <u>72.86</u> |
| Ours | CSPDarkNet-53 | - | 82.77 | 86.35 | 77.13 | 72.29 | 78.24 | 63.47 | 73.33 | 74.11 |

Table 5. Acc-Cls (%) on C-RefCOCO+/g. We compare our model with a random choice, different confidence scores and a binary classifier. Best results are in **bold** and sub-optimal results are underlined.

| Method | C-RefCOCO | | | C-RefCOCO+ | | | C-RefCOCOg | |
|--------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | val | testA | testB | val | testA | testB | val | test |
| Random | 51.12 | 49.64 | 51.95 | 53.33 | 51.74 | 49.07 | 49.89 | 50.03 |
| Conf. score (0.01) | 78.25 | 78.55 | 74.20 | 82.91 | 83.95 | 78.22 | <u>73.72</u> | <u>75.20</u> |
| Conf. score (0.1) | <u>86.25</u> | <u>86.05</u> | <u>82.67</u> | <u>87.88</u> | <u>87.62</u> | <u>84.15</u> | 62.08 | 61.81 |
| Conf. score (0.5) | 85.23 | 84.94 | 80.69 | 82.68 | 85.08 | 80.79 | 52.12 | 52.29 |
| Binary classifier | 93.05 | 92.31 | 91.69 | 91.98 | 91.09 | 89.21 | 89.45 | 89.08 |

while the weight parameters γ_{cf} and γ_{cl} are set to 2.0. The parameter α in DAF module is set to 0.25 for counterfactual head and 1.0 for regression head.

3.4. Baselines

We compare our model with state-of-the-art vision-language pretrained models and one-stage REC baselines. **The VLP models** include MDETR [17], OFA [43] and m-PLUG [21]. **One-stage baselines** include anchor-based models, such as FAOA [46], ReSC [47], MCN [28], RealGIN [51], LG-FPN [39] and SimREC [29] and anchor-free models, such as PFOS [38], VGTR [9] and TransVG [7].

3.5. Main Results

Table 4 reports Acc-Box of our model and other models on RefCOCO+/g. We observe that our model outperforms the prevailing one-stage REC models, especially the baseline SimREC [29]. Without being pre-trained on large-scale vision-language datasets, our model achieves strong competitive performance on the traditional REC task. It also implies that the training of dual tasks, i.e., localization and counterfactual classification can promote each other.

Table 6. Acc-Cf (%) of our model on C-RefCOCO+/g.

| Model | C-RefCOCO | | | C-RefCOCO+ | | | C-RefCOCOg | |
|-------|-----------|-------|-------|------------|-------|-------|------------|-------|
| | val | testA | testB | val | testA | testB | val | test |
| Ours | 86.32 | 86.99 | 82.44 | 82.31 | 82.99 | 74.38 | 78.18 | 78.27 |

Table 5 reports Acc-Cls of our model on constructed C-REC datasets, i.e., C-RefCOCO+/g. We set our baselines by a random choice and by confidence scores. We set three thresholds, including 0.01, 0.1, and 0.5 for confidence scores in regression head to directly predict counterfactual labels. The optimal thresholds obtain a performance gain from 23.83% to 36.41% compared to the random choice. However, the relatively low accuracy on C-RefCOCOg indicates that the method of thresholding confidence scores lacks of generalization when addressing the long and complex queries. In contrast, the binary classifier averagely obtains around 90% accuracy on C-RefCOCO+/g, surpassing the baseline methods by a margin up to 15.73%. This shows the necessity of a well-designed counterfactual prediction head on fine-grained samples.

Table 6 reports Acc-Cf of our model on constructed datasets C-RefCOCO+/g. As a combination of Acc-Box and Acc-Cls, Acc-Cf indicates the overall performance of a C-REC model. The high accuracy on all three datasets shows the strong performance of our model as a solid baseline for C-REC task.

3.6. Ablation Study

To investigate the effectiveness of different components in the model, we conduct ablation studies on our C-REC model. Specifically, we consider attribute features F_a , contrastive loss \mathcal{L}_{cl} , and counterfactual training (C-Train). Ta-

Table 7. Performance (%) on different settings of attribute features F_a , contrastive loss \mathcal{L}_{cl} , and counterfactual training (C-Train).

| F_a | \mathcal{L}_{cl} | C-Train | Acc-Box | Acc-Cls | Acc-Cf |
|-------|--------------------|---------|--------------|--------------|--------------|
| - | - | - | 75.23 | 50.00 | 37.62 |
| ✓ | - | - | 78.05 | 50.00 | 39.03 |
| - | ✓ | - | 77.96 | 50.00 | 38.98 |
| - | - | ✓ | 76.94 | 86.74 | 78.05 |
| ✓ | - | ✓ | 78.18 | 90.72 | 81.06 |
| ✓ | ✓ | ✓ | 78.28 | 91.69 | 82.44 |

Table 8. Acc-Cf (%) with different cross-modal fusion methods on C-RefCOCO.

| Method | val | testA | testB |
|--|--------------|--------------|--------------|
| Baseline (F_{vt}) | 83.72 | 84.83 | 79.29 |
| Serial fusion ($F_a \rightarrow F_{vt}$) | 82.04 | 82.14 | 78.35 |
| Serial fusion ($F_t \rightarrow F_{va}$) | 83.17 | 84.26 | 79.34 |
| Parallel fusion ($F_{va} + F_{vt}$) | 86.32 | 86.99 | 82.44 |

ble 7 reports the performance of our model under different settings on C-RefCOCO testB split. Note that here Acc-Box is only calculated on the positive samples in C-RefCOCO. Therefore, it is basically the box accuracy on a subset of RefCOCO. Methods without C-Train are only trained on positive samples. Consequently, they tend to predict $C = 1$ for all inputs and their Acc-Cls are fixed to 50%. Furthermore, we observe that the combination of three components brings the most significant gain of 44.82% on Acc-Cf. The results show that these designs do enhance the query-sensitive ability of our model.

Next, to investigate the effectiveness of our fusion method, we conduct different methods. Specifically, the baseline method is fusing the visual and textual features (F_{vt}) without additional guidance from attribute words. The other two baselines are serial methods, according to the phase of F_a incorporating into visual-textual feature fusion, i.e., calculating F_{vt} first or F_{va} first. The fourth method is parallel fusion method. The experimental results are reported in Table 8. Both of the serial methods show a slight performance decrease. In contrast, the parallel fusion method shows a balanced highest performance. This indicates that the parallel fusion method can better preserve the alignment between image and text modalities.

Furthermore, to evaluate the parameters’ impact on the model’s performance, we conduct experiments on the hyper-parameter α in DAF module and the temperature parameter τ in contrastive loss. **First**, the experimental results on the hyper-parameter α are shown in Figure 5. Here, we set different values of α for two subtasks, counterfactual detection and regression. We observe that the optimal values for these two subtasks are 0.25 (α_{cf}) and 1.00 (α_{reg}), respectively. The result shows that local attribute information is more important for detecting the counterfactual polarity,

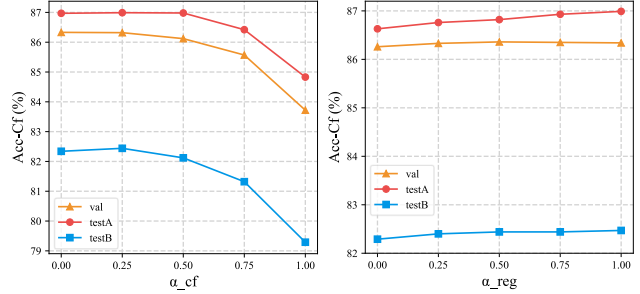


Figure 5. Acc-Cf (%) with different settings of the weight parameter α in DAF module on C-RefCOCO.

Table 9. Acc-Cf (%) with different settings of the temperature parameter τ in contrastive loss on C-RefCOCO.

| τ | val | testA | testB |
|--------|--------------|--------------|--------------|
| 0.05 | 85.43 | 86.42 | 81.78 |
| 0.1 | 85.47 | 86.25 | 81.39 |
| 0.2 | 86.32 | 86.99 | 82.44 |
| 0.5 | 84.83 | 86.68 | 80.90 |

while global context information is more important for localizing visual target referents. **Second**, we set different values of τ to train our model. The experimental results on the temperature parameter τ in contrastive loss are shown in Table 9. We observe that our model achieves the best performance when the parameter τ equals to 0.2. This indicates a higher or lower temperature parameter could weaken the feature fusion in latent feature space.

3.7. Qualitative Analysis

To intuitively show the performance of our model, we visualize a few qualitative examples in Figure 6. The first row shows the localization performance of our model and SimREC [29] on RefCOCO. Our model successfully predicts the counterfactual labels of the normal samples as positive and provides accurate localization. Even in the failure case “blue car”, bounding box predicted by our model has a higher IoU than that of SimREC. This shows a satisfactory performance of our model on traditional REC task. The other two rows show some counterfactual samples, covering all seven pre-defined attributes (see Table 1). Our model successfully identifies the mismatched attributes in most negative queries, including head noun, color, relative locations, and generic attribute. However, our model fails to focus on the size attribute “small” in the first failure case. In addition, in the second failure case, our model excessively attends to the absolute location “center” and produces the localization of “woman in the center”. These mistakes indicate that complex queries with multiple attributes can confuse our model to less attend to the exact counterfactual one.



Figure 6. Qualitative results of our C-REC model. At the top, we show some REC predictions by our model and SimREC on RefCOCO. At the bottom two rows, we show some C-REC predictions by our model, covering seven attributes. In all images, the ground truth boxes are marked in orange, the bounding boxes of SimREC in green and those of our model in blue.

4. Related Work

4.1. Counterfactual REC

There are several approaches that have studied on C-REC problem. SCRE [6] originally addresses the problem of wrong expressions in REC. The correctness of expressions is determined by the number of matched pairs of words and visual entities. MTG [10] takes a modular design of three components and outputs the logical union of masks from three segmentation models. These two methods take C-REC as a matching task based on logical rules. FVG [19] grounds the counterfactual queries to the pseudo regions added to the images. Moreover, other methods adopt a binary classifier to extend existing models. ReLA [24] addresses the no-target problem by a relationship-modeling framework. IRVG [22] handles the false-alarm issue by an iterative robust visual grounding framework. However, all of previous works do not investigate fine-grained counterfactual queries on various attributes.

4.2. Vision-language Counterfactuals Generation

In vision-language tasks such as REC and VQA, generating effective counterfactual image-text pairs is a necessity for evaluating counterfactual resilient models. A straightforward scheme is adopted in [6, 22]. They randomly match image-text pairs within existing datasets, leading to a relatively coarse quality. In addition, the idea of using generative adversarial networks (GANs) to produce counterfactual

images is proposed [2, 33]. CSS [4] generates both counterfactual images and questions by masking critical objects in images or keywords in texts. To obtain high-quality counterfactual queries, manual modifications on keywords in queries [6, 19] and manual re-annotations [24] are adopted. However, these schemes are labor-consuming and hard to transfer without explicit standards. In contrast, we propose a method based on language models to generate effective counterfactual text queries in a labor-free way.

5. Conclusion and Future Work

In this paper, we revisit the counterfactual problem in REC from a deep perspective of fine-grained attributes. To this end, we propose a CSG method to construct fine-grained C-REC datasets in a labor-free way. Furthermore, we propose a C-REC framework to detect the counterfactual polarity and simultaneously to locate the target referents. In addition, we incorporate a DAF module and contrastive learning to enhance counterfactual perception. Experimental results demonstrate that our C-REC model obtains promising performance on various datasets. In the future, how to generate counterfactual images to improve the diversity of our C-REC datasets is an interesting problem.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 6207603 and High-Performance Computing Platform of BUPT.

References

- [1] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2): 4606–4613, 2022. 1
- [2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020. 8
- [3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1
- [4] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809, 2020. 8
- [5] 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*, pages 1036–1044, 2021. 1
- [6] Enjie Cui, Jianming Wang, Jiayu Liang, and Guanghao Jin. Selective comprehension for referring expression by prebuilt entity dictionary with modular networks. In *Knowledge Management and Acquisition for Intelligent Systems: 15th Pacific Rim Knowledge Acquisition Workshop, PKAW 2018, Nanjing, China, August 28-29, 2018, Proceedings 15*, pages 211–220. Springer, 2018. 1, 8
- [7] 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. 1, 6
- [8] 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. 3
- [9] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 6
- [10] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019. 1, 8
- [11] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1
- [12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1
- [13] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 3
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3
- [16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 1
- [17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6
- [18] 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
- [19] Yongmin Kim, Chenhui Chu, and Sadao Kurohashi. Flexible visual grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 285–299, 2022. 8
- [20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [21] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 6
- [22] Menghao Li, Chunlei Wang, Wenquan Feng, Shuchang Lyu, Guangliang Cheng, Xiangtai Li, Binghao Liu, and Qi Zhao. Iterative robust visual grounding with masked reference based centerpoint supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4651–4656, 2023. 2, 8
- [23] 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 *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5
- [24] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 2, 8

- [25] 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. 1
- [26] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019. 1
- [27] Mingcong Lu, Ruifan Li, Fangxiang Feng, Zhanyu Ma, and Xiaojie Wang. Lgr-net: Language guided reasoning network for referring expression comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 1
- [28] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 6
- [29] Gen Luo, Yiyi Zhou, Jiamu Sun, Shubin Huang, Xiaoshuai Sun, Qixiang Ye, Yongjian Wu, and Rongrong Ji. What goes beyond multi-modal fusion in one-stage referring expression comprehension: An empirical study. *arXiv preprint arXiv:2204.07913*, 2022. 6, 7
- [30] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. 2, 3
- [31] 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. 5
- [32] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 5
- [33] Jingjing Pan, Yash Goyal, and Stefan Lee. Question-conditioned counterfactual image generation for vqa. *arXiv preprint arXiv:1911.06352*, 2019. 8
- [34] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1
- [36] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020. 1
- [37] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 1
- [38] Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, 2022. 6
- [39] Wei Suo, Mengyang Sun, Peng Wang, Yanning Zhang, and Qi Wu. Rethinking and improving feature pyramids for one-stage referring expression comprehension. *IEEE Transactions on Image Processing*, 32:854–864, 2022. 6
- [40] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 3, 5
- [41] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 1
- [42] Peng Wang, Qi Wu, Jiwei 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. 1
- [43] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 6
- [44] Nevan Wichers, Dilek Hakkani-Tür, and Jindong Chen. Resolving referring expressions in images with labeled elements. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 800–806. IEEE, 2018. 1
- [45] Qi Wu, Chunhua Shen, Anton Van Den Hengel, Peng Wang, and Anthony Dick. Image captioning and visual question answering based on attributes and their related external knowledge. *arXiv preprint arXiv:1603.02814*, 2, 2016. 1
- [46] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1, 6
- [47] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 6
- [48] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15502–15512, 2022. 1

- [49] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. [1](#)
- [50] 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. [1](#)
- [51] Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. [1](#), [3](#), [6](#)