Improving Referring Expression Grounding with Cross-modal Attention-guided Erasing

Xihui Liu[1] Zihao Wang[2] Jing Shao[2] Xiaogang Wang[1] Hongsheng Li[1]
[1]The Chinese University of Hong Kong [2] SenseTime Research
{xihuiliu, xgwang, hsli}@ee.cuhk.edu.hk {wangzihao, shaojing}@sensetime.com

Abstract

Referring expression grounding aims at locating certain objects or persons in an image with a referring expression, where the key challenge is to comprehend and align various types of information from visual and textual domain, such as visual attributes, location and interactions with surrounding regions. Although the attention mechanism has been successfully applied for cross-modal alignments, previous attention models focus on only the most dominant features of both modalities, and neglect the fact that there could be multiple comprehensive textual-visual correspondences between images and referring expressions. To tackle this issue, we design a novel cross-modal attention-guided erasing approach, where we discard the most dominant information from either textual or visual domains to generate difficult training samples online, and to drive the model to discover complementary textual-visual correspondences. Extensive experiments demonstrate the effectiveness of our proposed method, which achieves state-of-the-art performance on three referring expression grounding datasets.

1. Introduction

The goal of referring expression grounding [13, 39, 22] is to locate objects or persons in an image referred by natural language descriptions. Although much progress has been made in bridging vision and language [5, 32, 25, 37, 6, 2, 18], grounding referring expressions remains challenging because it requires a comprehensive understanding of complex language semantics and various types of visual information, such as objects, attributes, and relationships between regions.

Referring expression grounding is naturally formulated as an object retrieval task, where we retrieve a region that best matches the referring expression from a set of region proposals. Generally, it is difficult to trivially associate phrases and image regions in the embedding space where features are separately extracted from each modality (*i.e.*, vision and language). Previous methods [38, 10] proposed modular networks to handle expressions with different types of information. Another line of research explored attention mechanism, which mines crucial cues of both modalities [38, 4, 43]. By concentrating on the most important aspects in both modalities, the model with attention mechanism is able to learn better correspondences between words/phrases and visual regions, thus benefits the alignment between vision and language.

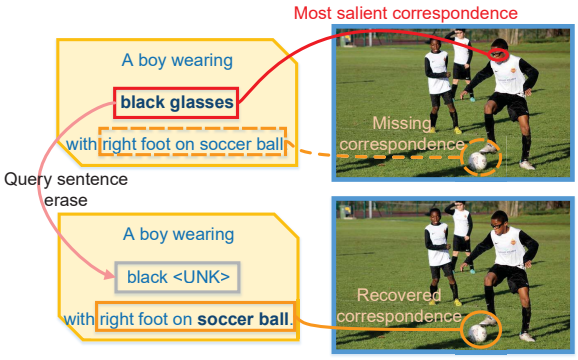

Figure 1. Query sentence erasing as an example of our cross-modal attention-guided erasing. The first row shows the original query-region pair, and the second row shows the pair with erased query.

However, a common problem of deep neural networks is that it tends to capture only the most discriminative information to satisfy the training constraints, ignoring other rich complementary information [42, 34]. This issue becomes more severe when considering attention models for referring expression grounding. By attending to both the referring expression and the image, the attention model is inclined to capturing the most dominant alignment between the two modalities, while neglecting other possible cross-modal correspondences. A referring expression usually describe an object from more than one perspectives, such as visual attributes, actions, and interactions with context objects, which cannot be fully explored by concentrating on only the most significant phrase-region pair. For example, people describe the image in Fig. 1 as "A boy wearing black glasses with right foot on soccer ball". We observe that the

model gives most attention on word "glasses", while ignoring other information like "soccer ball". As a result, the model can achieve a high matching score as long as it is able to recognize "glasses", and would fail to learn the visual features associated with the words "soccer ball". We argue that such limitations cause two problems: (1) it prevents the model from making full use of latent correspondences between training pairs. (2) A model trained in this way could overly rely on specific words or visual concepts and could be biased towards frequently observed evidences. Although some works on the recurrent or stacked attention [43, 4] perform multiple steps of attention to focus on multiple cues, they have no direct supervision on attention weights at each step and thus cannot guarantee that the models would learn complementary alignments rather than always focusing on similar information.

Inspired by previous works [29, 34] where they erase discovered regions to find complementary object regions, we design an innovative cross-modal erasing scheme to fully discover comprehensive latent correspondences between textual and visual semantics. Our cross-modal erasing approach erases the most dominant visual or textual information with high attention weights to generate difficult training samples online, so as to drive the model to look for complementary evidences besides the most dominant ones. Our approach utilizes the erased images with original queries, or erased queries with original images to form hard training pairs, and does not increase inference complexity. Furthermore, we take the interaction between image and referring expression into account, and use information from both self modality and the other modality as cues for selecting the most dominant information to erase. In particular, we leverage three types of erasing: (1) **Image-aware query sentence erasing**, where we use visual information as cues to obtain word-level attention weights, and replace the word with high attention weights with an "unknown" token. (2) **Sentence-aware subject region erasing**, where the spatial attention over subject region is derived based on both visual features and query information, and we erase the spatial features with the highest attention weights. (3) **Sentence-aware context object erasing**, where we erase a dominant context region, based on the sentence-aware object-level attention weights over context objects. Note that (2) and (3) are two complementary approaches for sentence-aware visual erasing. With training samples generated online by the *erasing* operation, the model cannot access the *most* dominant information, and is forced to further discover complementary textual-visual correspondences previously ignored.

To summarize, we introduce a novel cross-modal attention-guided erasing approach on both textual and visual domains, to encourage the model to discover comprehensive latent textual-visual alignments for referring expression grounding. To the best of our knowledge, this is the first work to consider erasing in both textual and visual domains to learn better cross-modal correspondences. To validate the effectiveness of our proposed approach, we conduct experiments on three referring expression datasets, and achieve state-of-the-art performance.

2. Related Work

Referring expression grounding. Referring expression grounding, also known as referring expression comprehension, is often formulated as an object retrieval task [11, 26]. [39, 23, 41] explored context information in images, and [31] proposed multi-step reasoning by multi-hop Feature-wise Linear Modulation. Hu *et al*. [10] proposed compositional modular networks, composed of a localization module and a relationship module, to identify subjects, objects and their relationships. Subsequent work by Yu *et al*. [38] built MattNet, which decomposes cross-modal reasoning into subject, location and relationship modules, and utilizes language-based attention and visual attention to focus on relevant components. [28, 22, 21, 40, 17] considered referring expression generation and grounding as inverse tasks, by either using one task as a guidance to train another, or jointly training both tasks. Our work is built upon MattNet, and encourages the model to explore complementary cross-modal alignments by cross-modal erasing.

Cross-modal Attention. Attention mechanism, which enables the model to select informative features, has been proven effective by previous works [35, 20, 3, 1, 36, 25, 14, 24, 16, 19]. In referring expression grounding, Deng *et al*. [4] proposed A-ATT to circularly accumulate attention for images, queries, and objects. Zhuang *et al*. [43] proposed parallel attention network with recurrent attention to global visual content and object candidates. To prevent the attention models from over-concentrating on the most dominant correspondences, we propose attention-guided erasing which generates difficult training samples on-the-fly, to discover complementary cross-modal alignments.

Adversarial erasing in visual Domain. Previous works has explored erasing image regions for object detection [33], person re-identification [12], weakly supervised detection [29, 9] and semantic segmentation [34]. Wang *et al*. [33] proposed to train an adversarial network that generates training samples with occlusions and deformations for training robust detector. Wei *et al*. [34] and Zhang *et al*. [42] proposed adversarial erasing for weakly supervised detection and segmentation, which drives the network to discover new and complementary regions by erasing the currently mined regions.

Different from previous works which only erase in visual domain, we take a step further towards cross-modal erasing in both images and sentences. More importantly, our approach only erases to create new training samples in the training phase, and does not increase inference complexity.

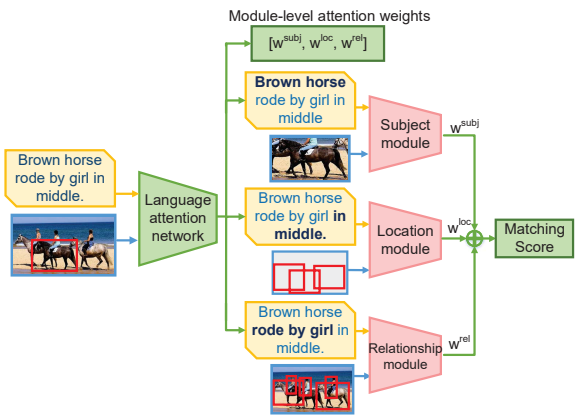

Figure 2. Illustration of our backbone model. The language attention network takes images and sentences as inputs, and outputs module-level attention weights and word-level attention weights for each module. The three visual modules calculate matching scores for subject, location and relationship, respectively. The final score is the weighted average of the three matching scores.

3. Cross-modal Attention-guided Erasing

Our cross-modal attention-guided erasing approach erases the most dominant information based on attention weights as importance indicators, to generate hard training samples, which drives the model to discover complementary evidences besides the most dominant ones. This approach is independent of the backbone architecture, and can be applied to any attention-based structures without introducing extra model parameters or inference complexity. In our experiments, we adopt the modular design of MattNet [38] as our backbone, because of its capability to handle different types of information in referring expressions.

3.1. Problem Formulation and Background

We formulate referring expression grounding as a retrieval problem: given an image I, a query sentence Q, and a set of region proposals $\mathcal{R} = \{R_i\}$ extracted from the image, we aim to compute a matching score between each region proposal R_i and the query Q, and the proposal with the highest matching score is chosen as the target object. For each region proposal R_i, its regional visual features together with context object features are denoted as O_i.

In MattNet [38], there is a language attention network and three visual modules, namely subject module, location module and relationship module. The language attention network takes the query Q as input, and outputs attention weights $\{w^{subj}, w^{loc}, w^{rel}\}$ and query embeddings for each module $[\mathbf{q}^{subj}, \mathbf{q}^{loc}, \mathbf{q}^{rel}]$. Each module calculates a matching score by dot product between the corresponding query embedding and visual or location features. The scores from three modules are fused according to the module-level attention weights $\{w^{subj}, w^{loc}, w^{rel}\}$. For positive candidate object and query pair (O_i, Q_i) and negative pairs (O_i, Q_j), (O_j, Q_i), the ranking loss is minimized during training:

$$L_{rank} = \sum_i ([m - s(O_i, Q_i) + s(O_i, Q_j)]_+ + [m - s(O_i, Q_i) + s(O_j, Q_i)]_+), \quad (1)$$

where $s(x, y)$ denotes the matching score between x and y, $[x]_+ = \max(x, 0)$, and m is the margin for ranking loss.

We adopt the modular structure of MattNet [38] and make some changes to the design of each module, which will be illustrated in Sec 3.3 to 3.5. The structure of our backbone is shown in Fig 2.

3.2. Overview of Attention-guided Erasing

By cross-modal erasing in both textual and visual domains to generate challenging training samples, we aim to discover complementary textual-visual alignments. (1) For **query sentence erasing**, we replace key words in the queries with the "unknown" token, and denote the erased referring expression as Q^*. (2) For **visual erasing**, we first select which visual module to erase based on the modular attention weights. Specifically, we sample a module according to the distribution defined by the module-level attention weights $M^s \sim \text{Multinomial}(3, [w_{subj}, w_{loc}, w_{rel}])$, and perform erasing on the inputs of the sampled module. For subject module which processes visual information of candidate objects, we perform **subject region erasing** on feature maps. For location and relationship modules which encode location or visual features of multiple context regions, we apply **context object erasing** to discard features of a context object. The erased features by either subject region erasing or context object erasing is denoted as O^*.

Given the erased query sentences or visual features, we replace the original samples with the erased ones in the loss function. Specifically, we force the erased visual features to match better with its corresponding queries than non-corresponding queries, and force the erased queries to match better with its corresponding visual features than non-corresponding ones, with the following erasing loss,

$$L_{erase} = \sum_i ([m - s(O_i^*, Q_i) + s(O_i^*, Q_j)]_+ + [m - s(O_i, Q_i^*) + s(O_j, Q_i^*)]_+). \quad (2)$$

where the first term forces matching between the erased visual features and original queries, and the second term forces matching between the erased queries and original visual features. We use a mixture of original and erased pairs in each mini-batch, and the overall loss is defined as,

$$L = L_{erase} + L_{rank}. \quad (3)$$

In the following, we discuss how to perform the three types of cross-modal attention-guided erasing, respectively.

3.3. Image-aware Query Sentence Erasing

People tend to describe a target object from multiple perspectives, but the model only focuses on the most dominant

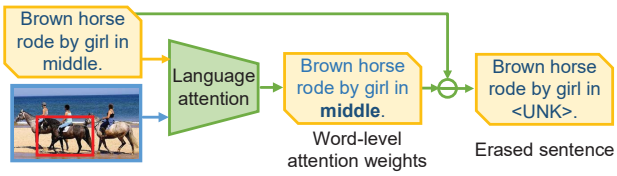

Figure 3. Image-aware query sentence erasing.

words, and neglects other words which may also imply rich alignments with visual information. Hence, we introduce erased queries into training to forbid the model from looking at only the most dominant word, so as to drive it to learn complementary textual-visual correspondences.

Image-aware module-level and word-level attention. Given the query sentence and the image, our first goal is to generate (1) attention weights for the three modules $\{w^{subj}, w^{loc}, w^{rel}\}$, and (2) three sets of word-level attention weights $\{\alpha_t^{subj}\}_{t=1}^T$, $\{\alpha_t^{loc}\}_{t=1}^T$, $\{\alpha_t^{rel}\}_{t=1}^T$ for three modules, where T is the number of words in the sentence.

Generally, understanding a referring expression not only requires the textual information, but also needs the image content as a cue. Inspired by this intuition, we design an image-aware language attention network to estimate module-level and word-level attention weights. Specifically, we encode the whole image $\mathbf{I}_0$ into a feature vector $\mathbf{e}_0$ with a convolutional neural network, and then feed the image feature vector and word embeddings $\{\mathbf{e}_t\}_{t=1}^T$ into the Long Short Term Memory Networks (LSTM).

$$\mathbf{e}_0 = \text{CNN}(\mathbf{I}_0), \quad \mathbf{h}_t = \text{LSTM}(\mathbf{e}_t, \mathbf{h}_{t-1}). \tag{4}$$

We calculate the module-level and word-level attention weights based on the hidden states of the LSTM, and derive query embedding for each module accordingly,

$$w^m = \frac{\exp(\mathbf{f}_m^T \mathbf{h}_T)}{\sum_{i\in\Omega} \exp(\mathbf{f}_i^T \mathbf{h}_T)}, \quad m \in \Omega, \tag{5}$$

$$\alpha_t^m = \frac{\exp(\mathbf{g}_m^T \mathbf{h}_t)}{\sum_{i=1}^T \exp(\mathbf{g}_m^T \mathbf{h}_i)}, \quad \mathbf{q}^m = \sum_{t=1}^T \alpha_t^m \mathbf{e}_t, \tag{6}$$

where $\mathbf{f}_m$ and $\mathbf{g}_m$ are model parameters, $\Omega = \{\text{subj}, \text{loc}, \text{rel}\}$ represents the three modules, and w^m denotes the model-level attention weights. α_t^m denotes the attention weight for word t and module m, and q^m is the query embedding for module m.

Our approach exploits visual cues to derive module-level and word-level attention weights, which is the key difference from previous works [38, 10] with only self-attention.

Attention-guided Query Erasing. Aiming to generate training samples by erasing the most important words in order to encourage the model to look for other evidences, we first calculate the overall significance of each word based on the module-level and word-level attention weights,

$$\alpha_t = \sum_{m\in\Omega} w^m \alpha_t^m, \tag{7}$$

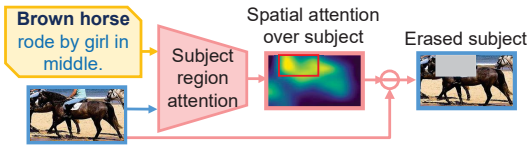

Figure 4. Sentence-aware subject region erasing.

where $\{\alpha_t\}_{t=1}^T$ denotes the image-aware overall attention weight for each word, which acts as an indicator of word importance. We sample a word to erase based on the distribution defined by overall word-level significance, $W^s \sim \text{Multinomial}(T, [\alpha_1, ..., \alpha_T])$.

Next, we consider in what way shall we eliminate the influence of this word. The most straightforward way is to directly remove it from the query sentence, but the sentence grammar would be broken in this way. For example, if we directly remove the word "chair" from the sentence "The gray office chair sitting behind a computer screen", the overall semantic meaning would be distorted and the model might have difficulty understanding it. In order to eliminate the influence of the erased word while preserving the sentence structure, we replace the target word with an "unknown" token, as shown in Fig. 3. In this way we obtain the erased query Q^*, which discards the semantic meaning of the erased word, but causes no difficulty for the model to understand the remaining words. The erased query Q_i^* and its original positive and negative image features O_i and O_j form new training sample pairs (O_i, Q_i^*) and (O_j, Q_i^*), and the we force textual-visual alignment between erased query sentences and original visual features by the ranking loss for erased query sentences (the second term in Eq.(2)).

3.4. Sentence-aware Subject Region Erasing

The subject module takes the feature map of a candidate region as input and outputs a feature vector. We create new training samples by erasing the most salient spatial features, to drive the model to discover complementary alignments.

Sentence-aware spatial attention. We follow previous works on cross-modal visual attention [38, 36, 4]. For a candidate region with its spatial features $\{\mathbf{v_j}\}_{j=1}^J$, where J is the number of spatial locations in the feature map, we concatenate the visual features at each location with the query embedding $\mathbf{q}^{subj}$ to calculate the spatial attention,

$$s_j = \mathbf{w}_2^s \tanh(\mathbf{W}_1^s[\mathbf{v}_j, \mathbf{q}^{subj}] + \mathbf{b}_1^s) + b_2^s, \tag{8}$$

$$\alpha_j^s = \frac{\exp(s_j)}{\sum_{i=1}^J \exp(s_i)}, \quad \tilde{\mathbf{v}}^{subj} = \sum_{j=1}^J \alpha_j^s \mathbf{v}_j, \tag{9}$$

where $\mathbf{W}_1^s$, $\mathbf{w}_2^s$, $\mathbf{b}_1^s$, b_2^s are model parameters, s_j is the unnormalized attention, α_j is the normalized spatial attention weights, and $\tilde{\mathbf{v}}^{subj}$ is the aggregated subject features.

Attention-guided subject region erasing. With conventional spatial attention, the model is inclined to focusing

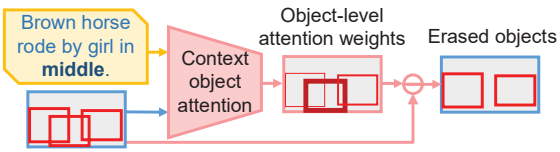

Figure 5. Sentence-aware object erasing for location module.

on only the most discriminative regions while neglecting other less salient regions. Such cases prevent the model from fully exploiting comprehensive textual-visual correspondences during training. So we erase salient features which are assigned greater attention weights to generate new training data, so as to drive the model to explore other spatial information and to learn complementary alignments.

In the feature map, spatially nearby features are correlated. Therefore, if we only erase features from separate locations, information of the erased features cannot be totally removed, since nearby pixels may also contain similar information. We therefore propose to erase a contiguous region of size $k \times k$ ($k = 3$ in our experiments) from the input feature map. In this way, the model is forced to look elsewhere for other evidences. Particularly, we calculate the accumulated attention weights of all possible regions in the feature map by a $k \times k$ sliding window, and mask the region with the highest accumulated attention weights (See Fig. 4 for illustration). The erased subject features together with original context object features are denoted as O_i^*. Similar to query sentence erasing, O_i^* is paired with original query sentences to form positive training samples (O_i^*, Q_i) and negative training samples (O_i^*, Q_j), and the ranking loss for visual erasing (the first term in Eq.(2)) is applied on the generated training sample pairs.

3.5. Sentence-aware Context Object Erasing

In referring expression grounding, supporting information from context objects (*i.e.* objects in the surrounding regions of the target object) is important to look for. For example, the expression "The umbrella held by woman wearing a blue shirt" requires an understanding of context region "woman wearing a blue shirt" and its relative location.

Sentence-aware attention over context objects. Sometimes multiple context regions are referred to in the sentence, *e.g.* "White sofa near two red sofas". So we formulate the location and relationship modules into a unified structure with sentence-aware attention, which considers multiple context objects, and attends to the most important ones.

For a set of context region features $\{\mathbf{c}_k^m\}_{k=1}^K$, where $m \in \{\text{loc}, \text{rel}\}$, and each $\mathbf{c}_k^m$ denotes the location or relationship feature of a context region proposal.[1] We derive object-level attention weights based on the concatenation of $\mathbf{c}_k^m$ and query embedding $\mathbf{q}^m$, and calculate the aggregated

[1]Details of context region selection and location and relationship feature extraction will be described in Sec 4.1.

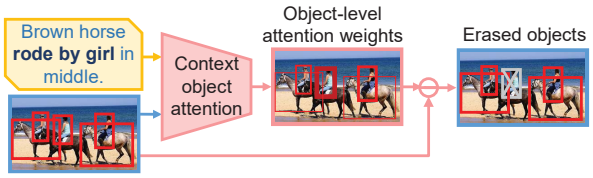

Figure 6. Sentence-aware object erasing for relationship module.

feature as the weighted sum of all object features,

$$s_k = \mathbf{w}_2^m \tanh(\mathbf{W}_1^m[\mathbf{c}_k^m, \mathbf{q}^m] + \mathbf{b}_1^m) + b_2^m, \tag{10}$$

$$a_k^m = \frac{\exp(s_k)}{\sum_{i=1}^K \exp(s_i)}, \quad \tilde{\mathbf{c}}^m = \sum_{k=1}^K a_k^m \mathbf{c}_k^m, \tag{11}$$

where $\mathbf{W}_1^m$, $\mathbf{w}_2^m$, $\mathbf{b}_1^m$, b_2^m are model parameters, s_k is the unnormalized scores, α_k^m is the normalized object-level attention weights, and $\tilde{\mathbf{c}}^m$ is the aggregated module features.

Our unified attention structure for location and relationship modules is different from MattNet [38]. In MattNet, the location module does not recognize different contributions of context regions, and the relationship module assumes only one context object contributes to recognizing the subject. In comparison, our model is able to deal with multiple context objects and attend to important ones, which is shown to be superior than MattNet in our experiments.

Attention-guided context object erasing. Sometimes the model may find the target region with the evidence from a certain context object, and hence do not need to care about other information. So we leverage attention-guided context object erasing to discard a salient context object, and use the erased contexts to form training samples, to encourage the model to look for subject or other supporting regions.

For both location and relationship modules, we obtain object-level attention weights over all considered objects $\{\alpha_k^m\}_{k=1}^K$ by sentence-aware context object attention. We sample a context object according to the attention weights $C^s \sim \text{Multinomial}(K, [\alpha_1, ..., \alpha_K])$, and discard C^s by replacing its features with zeros (see Fig. 5 and Fig. 6 for illustration). The erased context objects together with original subject features are denoted as O_i^*, which is paired with original query sentences to form positive training samples (O_i^*, Q_i) and negative training samples (O_i^*, Q_j), and the the ranking loss for visual erasing (the first term in Eq.(2)) is applied on the generated training sample pairs. The erased samples will drive the model to look for other context regions or subject visual features, and to discover complementary textual-visual alignments.

3.6. Theoretical Analysis

Back-propagation Perspective. We derive the gradients of attention models, and reveal that it emphasizes the gradients of the most salient features while suppresses the gradients of unimportant features. Such a conclusion validates the necessity of our proposed attention-guided erasing.

Consider the visual modality with features $\{\mathbf{f}_i\}_{i=1}^m$ and attention weights $\{\alpha_i\}_{i=1}^m$, and the textual modality with features $\{\mathbf{g}_j\}_{j=1}^n$ and attention weights $\{\beta_j\}_{j=1}^n$. The aggregated features are $\tilde{\mathbf{f}} = \sum_{i=1}^m \alpha_i \mathbf{f}_i$ and $\tilde{\mathbf{g}} = \sum_{j=1}^n \beta_j \mathbf{g}_j$, respectively. We calculate the cross-modal similarity as,

$$s = \tilde{\mathbf{f}}^\top \tilde{\mathbf{g}} = \big(\sum_{i=1}^m \alpha_i \mathbf{f}_i\big)^\top \big(\sum_{j=1}^n \beta_j \mathbf{g}_j\big) = \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j \mathbf{f}_i^\top \mathbf{g}_j \tag{12}$$

The gradient of s with respect to α_i, $\mathbf{f}_i$, β_j and $\mathbf{g}_j$ are

$$\frac{\partial s}{\partial \alpha_i} = \sum_{j=1}^n \beta_j \mathbf{f}_i^\top \mathbf{g}_j, \quad \frac{\partial s}{\partial \mathbf{f}_i} = \sum_{j=1}^n \alpha_i \beta_j \mathbf{g}_j, \tag{13}$$

$$\frac{\partial s}{\partial \beta_j} = \sum_{i=1}^m \alpha_i \mathbf{f}_i^\top \mathbf{g}_j, \quad \frac{\partial s}{\partial \mathbf{g}_j} = \sum_{i=1}^m \alpha_i \beta_j \mathbf{f}_i. \tag{14}$$

Suppose s is the matching score between the corresponding candidate region and the query sentence, and receives a positive gradient during back-propagation. If $\mathbf{f}_i$ and $\mathbf{g}_j$ are close to each other and $\mathbf{f}_i^\top \mathbf{g}_j > 0$, the attention weights α_i and β_j will receive positive gradients and be increased. On the contrary, if $\mathbf{f}_i^\top \mathbf{g}_j < 0$, both α_i and β_j will be tuned down. As a result, attention mechanism automatically learns importance of features without direct supervision.

On the other hand, if a word-region pair receives high attention α_i and β_j, the gradients with respect to $\mathbf{f}_i$ and $\mathbf{g}_j$ will be amplified, pushing $\mathbf{f}_i$ and $\mathbf{g}_j$ closer to each other to a large extent. While if α_i and β_j are small, the gradients will be suppressed, only pushing $\mathbf{f}_i$ and $\mathbf{g}_j$ slightly closer to each other. As a result, the model would learn large attention and good alignments only for the best aligned features, and updates inefficiently for other cross-modal alignments with low attention weights. Inspired by this analysis, our approach erases the best aligned features, forcing the model to give high attention weights to complementary cross-modal alignments, and to update those features efficiently.

Regularization Perspective. Our erasing mechanism can also be regarded as a regularization. The main difference from dropout [30] and dropblock [7] is that instead of randomly dropping features, we drop selectively. We erase salient information, as well as introducing randomness via sampling from the distributions defined by attention weights. The attention-guided erasing strategy is proven to be more effective than random erase in Sec. 4.5.

4. Experiments

4.1. Implementation Details

Visual feature representation. We follow MattNet [38] for feature representation of subject, location and relationship modules. We use faster R-CNN [27] with ResNet-101 [8] as backbone to extract image features, subject features and context object features. Specifically, we feed the whole image into faster R-CNN and obtain the feature map before ROI pooling as the whole image feature (used in Sec. 3.3). For each candidate object proposal, the 7×7 feature maps are extracted and fed into subject module (Sec. 3.4). For the location module, we encode the location features as the relative location offsets and relative areas to the candidate object $\delta l_{ij} = \big[\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}\big]$, as well as the position and relative area of the candidate object itself, *i.e.*, $l_i = \big[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{w \cdot h}{W \cdot H}\big]$. Attention and erasing for location module in Sec. 3.5 is performed over the location features of up-to-five surrounding same-category objects plus the candidate object itself. For relationship module, we use the concatenation of the average-pooled visual feature from the region proposal and relative position offsets and relative areas $\delta l_{ij} = \big[\frac{[\Delta x_{tl}]_{ij}}{w_i}, \frac{[\Delta y_{tl}]_{ij}}{h_i}, \frac{[\Delta x_{br}]_{ij}}{w_i}, \frac{[\Delta y_{br}]_{ij}}{h_i}, \frac{w_j h_j}{w_i h_i}\big]$ to represent relationship features of context objects. The attention and erasing on relationship module in Sec. 3.3 is performed over up-to-five surrounding objects.

Training Strategy. The faster R-CNN is trained on COCO training set, excluding samples from RefCOCO, RefCOCO+, and RefCOCOg's validation and test sets, and is fixed for extracting image and proposal features during training the grounding model. The model is trained with Adam optimizer [15] in two stages. We first pretrain the model by only original training samples with ranking loss $L = L_{rank}$ to obtain reasonable attention models for erasing. Then, we perform online erasing, and train the model with both original samples and erased samples generated online, with the loss function $L = L_{rank} + L_{erase}$.

4.2. Datasets and Evaluation Metrics

We conduct experiments on three referring expression datasets: RefCOCO (UNC RefExp) [39], RefCOCO+ [39], and RefCOCOg (Google RefExp) [22]. For RefCOCOg, we follow the data split in [23] to avoid the overlap of context information between different splits.

We adopt two settings for evaluation. In the first setting (denoted as ground-truth setting), the candidate regions are ground-truth bounding boxes, and a grounding is correct if the best-matching region is the same as the ground-truth. In the second setting (denoted as detection proposal setting), the model chooses the best-matching region from region proposals extracted by the object detection model, and a predicted region is correct if its intersection over union (IOU) with the ground-truth bounding box is greater than 0.5. Since our work focuses on textual-visual correspondence and comprehension of cross-modal information, rather than detection performance, we report results under both settings, and conduct analysis and ablation study with the first setting.

4.3. Results

Quantitative results. We show results of referring expression grounding compared with previous works under the

	test setting	RefCOCO			RefCOCO+			RefCOCOg		
		val	testA	testB	val	testA	testB	val*	val	test
MMI [22]	ground-truth	-	71.72	71.09	-	58.42	51.23	62.14	-	-
NegBag [23]	ground-truth	76.90	75.60	78.00	-	-	-	-	-	68.40
visdif+MMI [39]	ground-truth	-	73.98	76.59	-	59.17	55.62	64.02	-	-
Luo *et al*. [21]	ground-truth	-	74.04	73.43	-	60.26	55.03	65.36	-	-
CMN [10]	ground-truth	-	-		-	-	-	69.30	-	-
Speaker/visdif [39]	ground-truth	76.18	74.39	77.30	58.94	61.29	56.24	59.40	-	-
S-L-R [40]	ground-truth	79.56	78.95	80.22	62.26	64.60	59.62	72.63	71.65	71.92
VC [41]	ground-truth	-	78.98	82.39	-	62.56	62.90	73.98	-	-
Attr [17]	ground-truth	-	78.05	78.07	-	61.47	57.22	69.83	-	-
Accu-Att [4]	ground-truth	81.27	81.17	80.01	65.56	68.76	60.63	73.18	-	-
PLAN [43]	ground-truth	81.67	80.81	81.32	64.18	66.31	61.46	69.47	-	-
Multi-hop Film [31]	ground-truth	84.9	87.4	83.1	73.8	**78.7**	65.8	71.5	-	-
MattNet [38]	ground-truth	85.65	85.26	84.57	71.01	75.13	66.17	-	78.10	78.12
CM-Att	ground-truth	86.23	86.57	85.36	72.36	74.64	67.07	-	78.68	78.58
CM-Att-Erase	ground-truth	**87.47**	**88.12**	**86.32**	**73.74**	77.58	**68.85**	-	**80.23**	**80.37**
S-L-R [40]	det proposal	69.48	73.71	64.96	55.71	60.74	48.80	-	60.21	59.63
Luo [21]	det proposal	-	67.94	55.18	-	57.05	43.33	49.07	-	-
PLAN [43]	det proposal	-	75.31	65.52	-	61.34	50.86	58.03	-	-
MattNet [38]	det proposal	76.40	80.43	69.28	64.93	70.26	56.00	-	66.67	67.01
CM-Att	det proposal	76.76	82.16	70.32	66.42	72.58	57.23	-	67.32	67.55
CM-Att-Erase	det proposal	**78.35**	**83.14**	**71.32**	**68.09**	**73.65**	**58.03**	-	**67.99**	**68.67**

Table 1. Comparison with state-of-the-art referring expression grounding approaches on ground-truth regions and region proposals from detection model. For RefCOCO and RefCOCO+, testA is for grounding persons, and testB is for grounding objects.

ground-truth setting and detection proposal setting in Table 1. **CM-Att** denotes our model with cross-modal attention trained with only original training samples. **CM-Att-Erase** denotes our model with cross-modal attention trained with both original samples and erased samples generated by cross-modal attention-guided erasing. It is shown that the cross-modal attention model is already a strong baseline, and training with erased samples can further boost the performance. Our CM-Att-Erase model outperforms previous methods, without increasing inference complexity. It validates that with cross-modal erasing, the model is able to learn better textual-visual correspondences and is better at dealing with comprehensive grounding information.

Qualitative results. Fig. 7 shows qualitative results of our CM-Att-Erase model, compared with the CM-Att model. It is shown that our CM-Att-Erase model is better at handling complex information from both domains, especially for situations where multiple cues should be considered in order to ground the referring expressions. Take the second image in the first row as an example, our erasing model comprehends not only visual features associated with "dark blue flower pot" but also relationship with context object "pink flowers in it", while the model without erasing does not perform well for those cases.

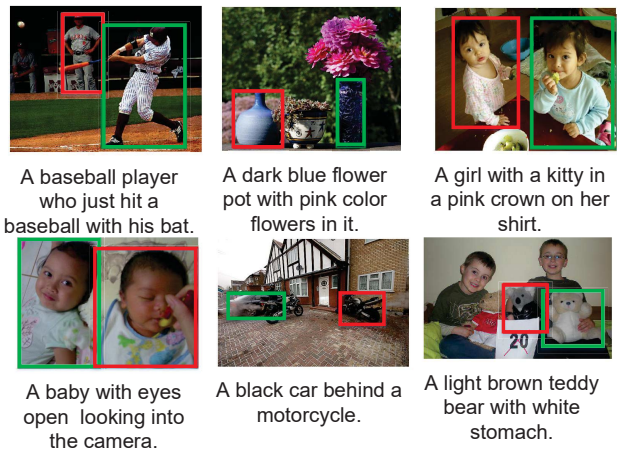

Figure 7. Qualitative results. Red bounding box denotes the grounding results of the CM-Att model, and green bounding box denotes grounding results of the CM-Att-Erase model.

4.4. Visualization of Attention and Erasing

We visualize the attention weights and erasing process in Fig. 8. It is shown that in the first image, the subject module gives high attention weights to the region corresponding to "black and white dress". However after erasing this region, the subject module attends on the action of this girl, encouraging the model to learn the correspondence between "playing tennis" and its corresponding visual features. The second line shows an example of query sentence erasing. By erasing the word "glasses" to obtain a new erased query as training sample, the model is driven to look for other information in the image, and it successfully identifies the alignment between "black phone" and the corresponding context object in the image.

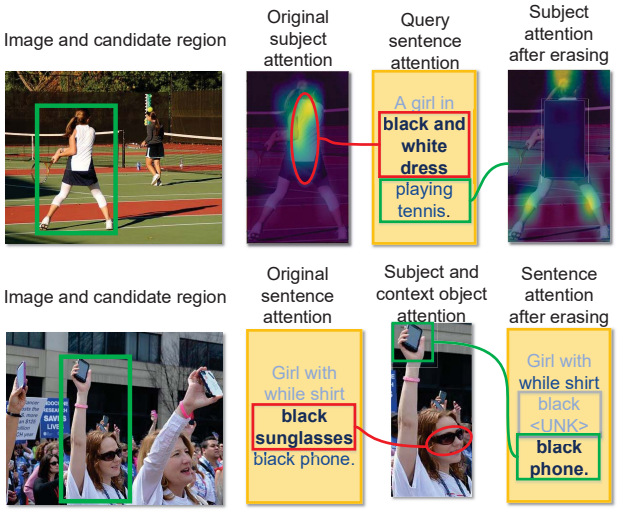

Figure 8. Visualization of attention weights before and after erasing. The first line shows an example of subject region erasing, and the second line shows an example of query sentence erasing.

		val	test
CM-Att-Erase (Our proposed approach)		80.23	80.37
Erasing methods	Random	79.08	79.05
	Adversarial network to erase	79.31	79.23
Effect of Cross-modal Erasing	Self-erasing	79.27	79.22
	Only textual erasing	79.21	79.55
	Only visual erasing	79.05	79.37
Iterative erasing		80.13	79.97
Erase during inference		79.25	79.56
Multiple steps of attention		79.31	78.49

Table 2. Ablation study results on RefCOCOg dataset.

4.5. Ablation Study

Erasing methods. Different choices of erasing methods were exploited by previous works. Other than our proposed attention-guided erasing, the most straightforward way is to randomly erase words or image regions without considering their importances [29]. Another choice is to train an adversarial network to select the most informative word or region to erase, which is used in [33]. We compare our attention-guided erasing approach with those methods, and results in Table 2 show that the attention-guided erasing performs better. Since attention weights are already good indicators of feature importance, leveraging attention as a guidance for erasing is more efficient, and the attention-guided erasing approach leads to little cost in model complexity, compared with applying a separate adversarial erasing network.

Effect of cross-modal erasing. We compare our cross-modal erasing approach with erasing based on self-attention weights, where we only utilize information within the same modality for generating attention weights and performing attention-guided erasing. We also experiment on only visual erasing or sentence erasing. Experimental results in Table 2 demonstrate the necessity of both visual erasing and query sentence erasing which are complementary to each other, and validate that our cross-modal attention-guided erasing is superior to self-attention-guided erasing without considering information from the other modality.

Iterative erasing. A possible extension is to iteratively perform multiple times of erasing similar to [34] to generate more challenging training samples progressively. However, results in Table 2 indicate that it is not suitable for this task. We observe that most referring expressions are quite short. Erasing more than one key words would significantly eliminate the semantic meaning of the sentence. Likewise, erasing the visual features for more than once would also make it impossible for the model to recognize the referred object.

Erasing during inference. Our model only leverages cross-modal erasing in the training phase and does not erase during inference. We try to erase key words or key regions during inference as well, and ensemble the matching scores of original samples and erased samples as the final score. But experiments suggest that it does not help the final performance. This is possibly because during training, the model have already learned to balance the weights of various features, and do not need to mask the dominant features to discover other alignments during inference.

Comparison with stacked attention. Leveraging multiple steps of attention also enables the model to attend to different features. However, those models do not pose direct constraints on learning complementary attention for different attention steps. We conduct experiments on stacked attention [36] to compare with our erasing approach. Experiments indicate that erasing performs better than stacked attention on this task, because by erasing we enforce stricter constraints of learning complementary alignments.

5. Conclusion and Future Work

We address the problem of comprehending and aligning various types of information for referring expression grounding. To prevent the model from over-concentrating on the most significant cues and drive the model to discover complementary textual-visual alignments, we design a cross-modal attention-guided erasing approach to generate hard training samples by discarding the most important information. The models achieve state-of-the-art performance on three referring expression grounding datasets, demonstrating the effectiveness of our approach.

Acknowledgements

This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616, CUHK14208417, CUHK14239816, and in part by CUHK Direct Grant.

References

[1] 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 vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[2] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–70, 2018.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017.

[4] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7746–7755, 2018.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.

[6] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven CH Hoi, and Xiaogang Wang. Question-guided hybrid convolution for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 469–485, 2018.

[7] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *arXiv preprint arXiv:1810.12890*, 2018.

[8] 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.

[9] Qibin Hou, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. *arXiv preprint arXiv:1810.09821*, 2018.

[10] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE, 2017.

[11] 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.

[12] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018.

[13] 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.

[14] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018.

[17] Jingyu Liu, Liang Wang, Ming-Hsuan Yang, et al. Referring expression generation and comprehension via attributes. In *Proceedings of CVPR*, 2017.

[18] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–354, 2018.

[19] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[20] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, 2017.

[21] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[22] 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.

[23] 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.

[24] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *arXiv preprint arXiv:1611.00471*, 2016.

[25] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. *arXiv preprint arXiv:1804.00775*, 2018.

[26] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. *arXiv preprint arXiv:1711.08389*, 2017.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[28] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[29] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[31] Florian Strub, Mathieu Seurin, Ethan Perez, Harm de Vries, Philippe Preux, Aaron Courville, Olivier Pietquin, et al. Visual reasoning with multi-hop feature modulation. *arXiv preprint arXiv:1808.04446*, 2018.

[32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015.

[33] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[34] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, 2017.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[36] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[37] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–338, 2018.

[38] 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 (CVPR)*, 2018.

[39] 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.

[40] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[41] 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.

[42] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas Huang. Adversarial complementary learning for weakly supervised object localization. In *IEEE CVPR*, 2018.

[43] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D Reid, and Anton van den Hengel. Parallel attention: a unified framework for visual object discovery through dialogs and queries. In *Proceedings of CVPR*, 2018.