Cross-Modal Relationship Inference for Grounding Referring Expressions

Sibei Yang[1] Guanbin Li[2] Yizhou Yu[1,3]

[1]The University of Hong Kong [2]Sun Yat-sen University [3]Deepwise AI Lab

sbyang9@hku.hk, liguanbin@mail.sysu.edu.cn, yizhouy@acm.org

Abstract

*Grounding referring expressions is a fundamental yet challenging task facilitating human-machine communication in the physical world. It locates the target object in an image on the basis of the comprehension of the relationships between referring natural language expressions and the image. A feasible solution for grounding referring expressions not only needs to extract all the necessary information (*i.e. *objects and the relationships among them) in both the image and referring expressions, but also compute and represent multimodal contexts from the extracted information. Unfortunately, existing work on grounding referring expressions cannot extract multi-order relationships from the referring expressions accurately and the contexts they obtain have discrepancies with the contexts described by referring expressions. In this paper, we propose a Cross-Modal Relationship Extractor (CMRE) to adaptively highlight objects and relationships, that have connections with a given expression, with a cross-modal attention mechanism, and represent the extracted information as a language-guided visual relation graph. In addition, we propose a Gated Graph Convolutional Network (GGCN) to compute multimodal semantic contexts by fusing information from different modes and propagating multimodal information in the structured relation graph. Experiments on various common benchmark datasets show that our Cross-Modal Relationship Inference Network, which consists of CMRE and GGCN, outperforms all existing state-of-the-art methods.*[1]

1. Introduction

Understanding natural languages and their relationship with visual information is the foundation in AI for bridging humans and machines in the physical world. This problem involves many challenging tasks, among which, grounding referring expressions [9, 19] is a fundamental one. Grounding referring expressions attempts to locate the target visual object in an image on the basis of comprehending the relationships between referring natural language expressions (e.g. "the man with glasses", "the dog near a car") and the image.

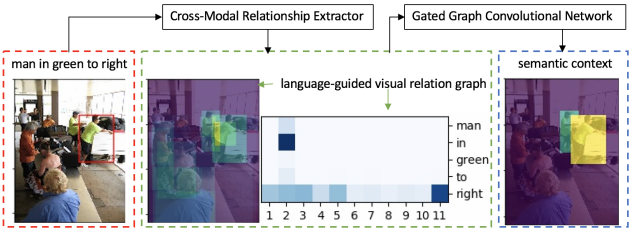

Figure 1. Cross-Modal Relationship Inference Network. Given an expression and image, Cross-Modal Relationship Extractor constructs the language-guided visual relation graph (the attention scores of proposals and edges' types are visualized inside green dashed box). The Gated Graph Convolutional Network captures semantic contexts and computes the matching score between the context of every proposal and the context of the expression (the matching scores of proposals are shown inside blue dashed box). Warmer color indicates higher scores of pixels and darker blue indicates higher scores of edges' types.

Grounding referring expressions is typically formulated as a task that identifies a proposal referring to the expressions from a set of proposals in an image [30]. Recent work combines Convolutional Neural Networks (CNN) [13] and Long Short-Term Memory Neural Networks (LSTM) [6] to process the multimodal information in images and referring expressions. CNNs are used for extracting visual features of single objects, global visual contexts [19, 24] and pairwise visual differences [15, 30, 31, 32], while LSTMs encode global language contexts [15, 18, 19, 31, 32] and language features of the decomposed phrases [7, 30, 36, **?**]. CNN cooperating with LSTM can also capture the context of object pairs [7, 22, 36]. However, such work cannot extract first-order relationships or multi-order relationships accurately from referring expressions, and the captured contexts also

[1]Code is available at https://github.com/sibeiyang/sgmn/tree/master/lib/cmrin_models. If you have any inquires, please feel free to contact Sibei Yang via sbyang@cs.hku.hk. Corresponding author is Guanbin Li. This work was partially supported by the Hong Kong PhD Fellowship, the National Natural Science Foundation of China under Grant No.61702565 and the Fundamental Research Funds for the Central Universities under Grant No.18lgpy63.

have discrepancies with the contexts described by referring expressions.

A feasible solution for grounding referring expressions needs to extract all the required information (*i.e*. objects and the relationships among them in the image and referring expressions) accurately for any given expression. However, as the expressions generated from the scene in an image are often unpredictable and flexible [19], the proposed model needs to extract information adaptively. For example, if the target is to locate "The man holding a balloon" in an image with two or more men, the required information from the natural language expression includes nouns ("man" and "balloon") and the word about relationship "holding"; on the other hand, the proposals for "man" and "balloon" and the visual relationship ('holding") linking them should be identified in the image. If the expression is more complicated, such as "The man on the left of the man holding a balloon", the additional relation information we need is "left". In this example, we need to recognize the second-order relationship between the target and the "balloon" through the other "man". Unfortunately, existing work either does not support relationship modeling or only considers first-order relationships among objects [7, 22, 36]. Theoretically, visual relation detectors [3, 17, 33] and natural language parsers can help achieve that goal by detecting the relational information in the image and parsing the expressions in the language mode. However, existing visual relation detectors cannot deliver satisfactory results for highly unrestricted scene compositions [36], and existing language parsers have adverse effects on the performance of grounding referring expressions due to their parsing errors [30].

Moreover, it is vital to represent the contextual information of referring expressions and target object proposals accurately and consistently because the context of an expression helps distinguish the target from other objects [22, 31, 36]. Nevertheless, existing methods for context modeling either cannot represent the contexts accurately or cannot achieve high-level consistency between both types of contexts mentioned above, and the reasons are given below. First, existing work on global language context modeling [15, 18, 19, 31, 32] and global visual context modeling [19, 24] introduces noisy information and makes it hard to match these two types of contexts. Second, pairwise visual differences computed in existing work [15, 30, 31, 32] can only represent instance-level visual differences among objects of the same category. Third, existing work on context modeling for object pairs [7, 22, 36] only considers first-order relationships but not multi-order relationships (e.g., they directly extract the context between the target "man" and "balloon" without considering the other "man" "holding the balloon"). In addition, multi-order relationships are actually structured information, and the context encoders adopted by existing work on grounding referring expressions are simply incapable of modeling them.

In order to overcome the aforementioned difficulties, we propose an end-to-end Cross-Modal Relationship Inference Network (CMRIN). CMRIN consists of two modules, *i.e*. the Cross-Modal Relationship Extractor (CMRE) and the Gated Graph Convolutional Network (GGCN). An example is illustrated in Figure 1. The CMRE extracts all the required information adaptively (*i.e*., nouns and relationship words from the expressions, and object proposals and their visual relationships from the image) for constructing a language-guided visual relation graph with cross-modal attention. First, CMRE constructs a spatial relation graph for the image. Second, it learns to classify the words in the expression into four types and further assign the words to the vertices and edges in the spatial relation graph. Finally, it constructs the language-guided visual relation graph from the normalized attention distribution of words over vertices and edges of the spatial relation graph. The GGCN fuses information from different modes and propagates the fused information in the language-guided visual relation graph to obtain the semantic context referred to by the expression. We have tested our proposed CMRIN on three common benchmark datasets, including RefCOCO [31], RefCOCO+ [31] and RefCOCOg [19]. Experimental results show that our proposed network outperforms all the other state-of-the-art methods.

In summary, this paper has the following contributions:

- Cross-Modal Relationship Extractor (CMRE) is proposed to convert the pair of input expression and image into a language-guided visual relation graph. For any given expression, CMRE highlights objects as well as relationships among them with a cross-modal attention mechanism.
- Gated Graph Convolutional Network (GGCN) is proposed to capture multimodal semantic contexts with multi-order relationships. GGCN fuses information from different modes and propagates fused information in the language-guided visual relation graph.
- CMRE and GGCN are integrated into Cross-Modal Relationship Inference Network (CMRIN), which outperforms all existing state-of-the-art methods on grounding referring expressions.

2. Related Work

2.1. Grounding Referring Expressions

Grounding referring expression and referring expression generation [19] are dual tasks. The latter generates an unambiguous text expression for a target object in an image, and the former selects the corresponding object according to the context in an image referred to by a text expression.

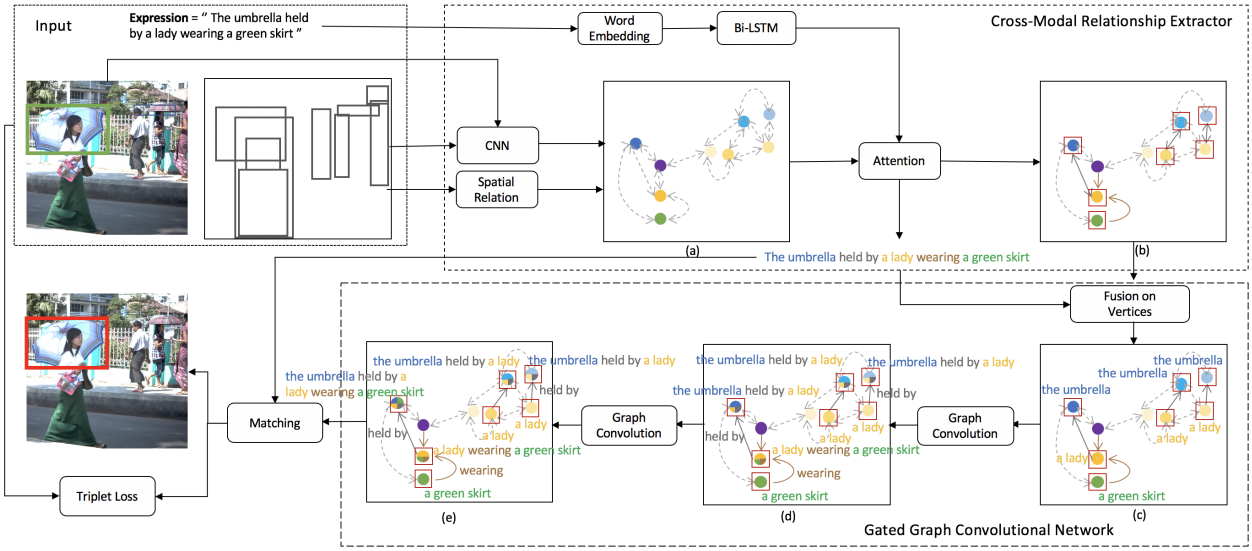

Figure 2. An overview of our Cross-Modal Relationship Inference Network for grounding referring expressions (better view in color). We use color to represent semantics, *i.e.* yellow means "person", green means "green shirt", blue means "umbrella", purple means "white T-shirt", brown means "wearing" and dark grey means "held by". It includes a Cross-Modal Relationship Extractor (CMRE) and a Gated Graph Convolutional Network (GGCN). First, CMRE constructs (a) a spatial relation graph from the visual features of object proposals and spatial relationships between proposals. Second, CMRE highlights the vertices (red bounding boxes) and edges (solid lines) to generate (b) a language-guided visual relation graph using cross-modal attention between words in the referring expression and the spatial relation graph's vertices and edges. Third, GGCN fuses the context of every word into the language-guided visual relation graph to obtain (c) a multimodal (language, visual and spatial information) relation graph. Fourth, GGCN captures (d) multimodal semantic contexts with first-order relationships by performing gated graph convolutional operations in the relation graph. By performing gated graph convolutional operations for multiple iterations, (e) semantic contexts with multi-order relationships can be computed. Finally, CMRIN calculates the matching score between the semantic context of every proposal and the global context of the referring expression.

To address grounding referring expression, some previous work [15, 18, 19, 32, 31] extracts visual object features from CNN and treats an expression as a whole to encode language feature through an LSTM. Among them, some methods [18, 19, 31] learn to maximize the posterior probability of the target object given the expression and the image, and the others [15, 32] model the joint probability of the target object and the expression directly. Different from the methods above, Context Modeling between Objects Network [22] inputs the concatenation of visual object representation, visual context representation and the word embedding to an LSTM model. Some recent methods [7, 30, 36] learn to decompose an expression into different components and compute the language-vision matching scores of each module for objects, others [37, 4] adopt co-attention mechanisms to build up the interactions between the expression and the objects in the image. Our Cross-Modal Relationship Extractor also learns to classify the words in the expression, but we treat the classified words as a guidance to highlight all the objects and their relationships described in the expression automatically to build the language-guided visual relation graph without extra visual relationships detection [3, 17, 33].

2.2. Context modeling

Context modeling has been utilized in many visual recognition tasks, *e.g.*, object detection [2, 27, 34] and semantic segmentation [21, 35]. Recently, Structure Inference Network [16] formulates the context modeling task as a graph structure inference problem [8, 11, 20]. Previous work on grounding referring expressions also attempts to capture contexts. Some work [19, 24] encodes the entire image as a visual context, but that global contextual information usually cannot match with the local context described by expression. Some work [15, 30, 31, 32] captures the visual difference between the objects belonging to the same category in an image, but the visual difference of the object's appearance is often insufficient to distinguish the target from other objects. Instead, the visual difference between the context including appearance and relationship is essential, *e.g.*, "Man holding a balloon", the necessary information to locate the "man" is not only the appearance of the "man" but the "holding" relation with the "balloon". Though there are some work [7, 22, 36] attempt to model the context from the relationship of object pairs, they only consider the context with first-order relationship between objects. Inspired by Graph Convolutional Network [11]

for classification, our Gated Graph Convolutional Network flexibly captures the context referring to the expression by message passing, and the context with multi-order relationships can be captured.

3. Cross-Modal Relationship Inference Network

Our proposed Cross-Modal Relationship Inference Network (CMRIN) relies on cross-modal relationships among objects and contexts captured in the multimodal relation graph to choose the target object proposal (in the input image) referred to by the input expression. First, CMRIN constructs a language-guided visual relation graph using the Cross-Modal Relationship Extractor. Second, it captures multimodal contexts from the relation graph using the Gated Graph Convolutional Network. Finally, a matching score is computed for each object proposal according to its multimodal context and the context of the input expression. The overall architecture of our CMRIN for grounding referring expressions is illustrated in Figure 2. In the rest of this section, we elaborate all the modules in this network.

3.1. Cross-Modal Relationship Extractor

The Cross-Modal Relationship Extractor (CMRE) adaptively constructs the language-guided visual relation graph according to any given pair of image and expression using a cross-modal attention mechanism. Our CMRE softly classifies the words in the expression into four types (*i.e.*, entity words, relation, absolute location, and unnecessary words) according to the context of every word. The context of the entire expression can be calculated from the context of each individual word. Meanwhile, a spatial relation graph of the image is constructed by linking object proposals in the image according to their size and locations. Next, CMRE generates the language-guided visual relation graph by highlighting the vertices and edges of the spatial relation graph. Highlighting is implemented as computing cross-modal attention between the words in the expression and the vertices and edges in the spatial relation graph.

3.1.1 Spatial Relation Graph

Exploring relative spatial relations among object proposals within an image is necessary for grounding referring expressions. On one hand, spatial information frequently occurs in referring expressions [30]; on the other hand, spatial relationship is an important aspect of visual relationship in images [3, 33]. We explore the spatial relationship between each pair of proposals according to their size and locations, which bears resemblance to the approach in [29].

For a given image I with K object proposals (bounding boxes), $O = \{o_i\}_{i=1}^{K}$, the location of each proposal o_i is denoted as $loc_i = (x_i, y_i, w_i, h_i)$, where (x_i, y_i) are the normalized coordinates of the center of proposal o_i, w_i and h_i are the normalized width and height respectively. The spatial feature $\mathbf{p}_i$ is defined as $\mathbf{p}_i = [x_i, y_i, w_i, h_i, w_i h_i]$. For any pair of proposals o_i and o_j, the spatial relationship r_{ij} between them is defined as follows. We compute the relative distance d_{ij}, relative angle θ_{ij} (*i.e.* the angle between the horizontal axis and vector $(x_i - x_j, y_i - y_j)$) and Intersection over Union u_{ij} between them. If o_i includes o_j, r_{ij} is set to "inside"; if o_i is covered by o_j, r_{ij} is set to "cover"; if none of the above two cases is true and u_{ij} is larger than 0.5, r_{ij} is set to "overlap"; otherwise, when the ratio between d_{ij} and the diagonal length of the image is larger than 0.5, r_{ij} is set to "no relationship". In the rest of the cases, r_{ij} is assigned to one of the following spatial relations, "right", "top right", "top", "top left", "left", "bottom left", "bottom" and "bottom right", according to the relative angle θ_{ij}. The details are shown in Figure 3.

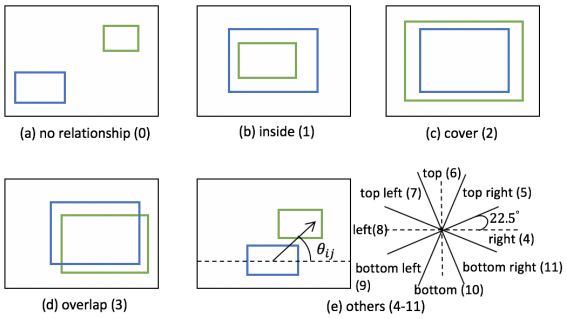

Figure 3. All types of spatial relations between proposal o_i (green box) and proposal o_j (blue box). The number following the relationship is the label index.

The directed spatial relation graph $G^s = (V, E, \mathbf{X}^s)$ is constructed from the set of object proposals O and the set of pairwise relationships $R = \{r_{ij}\}_{i,j=1}^{K}$, where $V = \{v_i\}_{i=1}^{K}$ is the set of vertices and vertex v_i corresponds to proposal o_i; $E = \{e_{ij}\}_{i,j=1}^{K}$ is the set of edges and e_{ij} is the index label of relationship r_{ij}; $\mathbf{X}^s = \{\mathbf{x}_i^s\}_{i=1}^{K}$ is the set of features at vertices and $\mathbf{x}_i^s \in \mathbb{R}^{D_x}$ is the visual feature of proposal o_i, and $\mathbf{x}_i^s$ is extracted using a pretrained CNN model. A valid index label of E ranges from 1 to $N_e = 11$ (the label of "no relationship" is 0).

3.1.2 Language Context

Inspired by the attention weighted sum of word vectors over different modules in [7, 36, 30], our CMRE defines attention distributions of words over the vertices and edges of the spatial relation graph G^s. In addition, different words in a referring expression may play different roles. For referring expressions, words can usually be classified into four types, *i.e.* entity, relation, absolute location and unnecessary words. By classifying the words into different types and distributing them over the vertices and edges of graph G^s, the

language context of every vertex and edge can be captured, and the global language context can also be obtained.

For a given expression $L = \{l_t\}_{t=1}^{T}$, CMRE uses a bi-directional LSTM [26] to encode the context of every word. The context of word l_t is the concatenation of its forward and backward hidden vectors, denoted as $\mathbf{h}_t \in \mathbb{R}^{D_h}$. The weight $\mathbf{m}_t$ of each type (*i.e.* entity, relation, absolute location and unnecessary word) for word l_t is defined as follows.

$$\mathbf{m}_t = \text{softmax}(\mathbf{W}_{l1}\sigma(\mathbf{W}_{l0}\mathbf{h}_t + \mathbf{b}_{l0}) + \mathbf{b}_{l1}), \quad (1)$$

where $\mathbf{W}_{l0} \in \mathbb{R}^{D_{l0}\times D_h}$, $\mathbf{b}_{l0} \in \mathbb{R}^{D_{l0}\times 1}$, $\mathbf{W}_{l1} \in \mathbb{R}^{4\times D_{l0}}$ and $\mathbf{b}_{l1} \in \mathbb{R}^{4\times 1}$ are learnable parameters, D_{l0} and D_h are hyper-parameters and σ is the activation function. The weight of entity, relation and absolute location are the first three elements of $\mathbf{m}_t$. The global language context $\mathbf{h}_g$ of graph G^s is calculated as $\mathbf{h}_g = \sum_{t=0}^{T}(\mathbf{m}_t^{(0)} + \mathbf{m}_t^{(1)} + \mathbf{m}_t^{(2)})\mathbf{h}_t$.

Next, on the basis of the word contexts $\mathbf{H} = \{\mathbf{h_t}\}_{t=1}^{T}$ and the entity weight of every word $\{\mathbf{m}_t^{(0)}\}_{t=1}^{T}$, a weighted normalized attention distribution over the vertices of graph G^s is defined as follows.

$$\begin{aligned}\alpha_{t,i} &= \mathbf{W}_n[\tanh(\mathbf{W}_v\mathbf{x}_i^s + \mathbf{W}_h\mathbf{h}_t)],\\ \lambda_{t,i} &= \mathbf{m}_t^{(0)}\frac{\exp(\alpha_{t,i})}{\sum_i^K \exp(\alpha_{t,i})},\end{aligned} \quad (2)$$

where $\mathbf{W}_n \in \mathbb{R}^{1\times D_n}$, $\mathbf{W}_v \in \mathbb{R}^{D_n\times D_x}$ and $\mathbf{W}_h \in \mathbb{R}^{D_n\times D_h}$ are transformation matrices and D_n is hyper-parameter. $\lambda_{t,i}$ is the weighted normalized attention, indicating the probability that word l_t refers to vertex v_i. The language context $\mathbf{c}_i$ at vertex v_i is calculated by aggregating all attention weighted word contexts.

$$\mathbf{c}_i = \sum_{t=1}^{T}\lambda_{t,i}\mathbf{h}_t. \quad (3)$$

3.1.3 Language-Guided Visual Relation Graph

Different object proposals and different relationships between proposals do not have equal contributions in solving grounding referring expressions. The proposals and relationships mentioned in the referring expression should be given more attention. Our CMRE highlights the vertices and edges of the spatial relation graph G^s, that have connections with the referring expression, to generate the language-guided visual relation graph G^v. The highlighting operation is implemented by designing a gate for each vertex and edge in graph G^s.

The gate p_i^v for vertex v_i is defined as the sum over the weighted probabilities that individual words in the expression refer to vertex v_i,

$$p_i^v = \sum_{t=1}^{T}\lambda_{t,i} \quad (4)$$

Each edge has its own type and the gates for edges are formulated as the gates for edges' types. The weighted normalized distribution of words over the edges of graph G^s is defined as follows.

$$\mathbf{w}_t^e = \text{softmax}(\mathbf{W}_{e1}\sigma(\mathbf{W}_{e0}\mathbf{h}_t + \mathbf{b}_{e0}) + \mathbf{b}_{e1})\mathbf{m}_t^{(1)}, \quad (5)$$

where $\mathbf{W}_{e0} \in \mathbb{R}^{D_{e0}\times D_h}$, $\mathbf{b}_{e0} \in \mathbb{R}^{D_{e0}\times 1}$, $\mathbf{W}_{e1} \in \mathbb{R}^{N_e\times D_{e0}}$ and $\mathbf{b}_{e1} \in \mathbb{R}^{N_e\times 1}$ are learnable parameters, and D_{e0} is hyper-parameter. $w_{t,j}^e$ is the j-th element of $\mathbf{w}_t^e$, which is the weighted probability of word l_t referring to edge type j. And the gate p_j^e for edges with type $j \in \{1, 2, ..N^e\}$ is the sum over all the weighted probabilities that individual words in the expression refer to edge type j,

$$p_j^e = \sum_{t=1}^{T} w_{t,j}^e. \quad (6)$$

The language-guided visual relation graph is defined as $G^v = (V, E, \mathbf{X}, P^v, P^e)$, where $P^v = \{p_i^v\}_{i=1}^{K}$, and $P^e = \{p_j^e\}_{j=1}^{N_e}$.

3.2. Multimodal Context Modeling

Our proposed Gated Graph Convolutional Network (GGCN) further fuses the language contexts into the language-guided visual relation graph to generate multimodal relation graph G^m, and computes a multimodal semantic context for every vertex by performing gated graph convolutional operations on the graph G^m.

3.2.1 Language-Vision Feature

As suggested by visual relationships detection [3, 33], the spatial locations together with the appearance features of objects are the key indicators of visual relationship, and the categories of objects is highly predictive of relationship. Our GGCN fuses the language context of every vertex into the language-guided visual relation graph G^v (G^v encodes the spatial relationships and appearance features of proposals) to generate multimodal relation graph G^m, which forms the basis for computing the semantic context of every vertex.

We define feature $\mathbf{x}_i^m$ at vertex v_i in G^m to be the concatenation of the visual feature $\mathbf{x}_i^s$ at vertex v_i in the language-guided visual relation graph and the language context $\mathbf{c}_i$ at vertex v_i, *i.e.* $\mathbf{x}_i^m = [\mathbf{x}_i^s, \mathbf{c}_i]$. The multimodal graph is defined as $G^m = (V, E, \mathbf{X}^m, P^v, P^e)$, where $\mathbf{X}^m = \{\mathbf{x}_i^m\}_{i=1}^{K}$.

3.2.2 Semantic Context Modeling

Multi-order relationships may exist in referring expressions. We obtain semantic contexts representing multi-order relationships through message passing. On one hand, semantic features are obtained by learning to fuse the spatial relations, visual features and language features. On the other

hand, contexts representing multi-order relationships are computed by propagating pairwise contexts in graph G^m.

Inspired by Graph Convolutional Network (GCN) for classification [11, 28], our GGCN adopts graph convolutional operations in multimodal relation graph G^m for computing semantic contexts. Different from GCN operating in unweighted graphs, GGCN operates in weighted directed graphs with extra gate operations. The n-th gated graph convolution operation at vertex v_i in graph $G^m = (V, E, \mathbf{X}^m, P^v, P^e)$ is defined as follows.

$$\begin{aligned}
\overrightarrow{\mathbf{x}}_i^{(n)} &= \sum_{e_{i,j}>0} p^e_{e_{i,j}} (\overrightarrow{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_j^{(n-1)} p_j^v + \mathbf{b}^{(n)}_{e_{i,j}}), \\
\overleftarrow{\mathbf{x}}_i^{(n)} &= \sum_{e_{j,i}>0} p^e_{e_{j,i}} (\overleftarrow{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_j^{(n-1)} p_j^v + \mathbf{b}^{(n)}_{e_{j,i}}), \\
\tilde{\mathbf{x}}_i^{(n)} &= \widetilde{\mathbf{W}}^{(n)} \hat{\mathbf{x}}_i^{(n-1)} + \widetilde{\mathbf{b}}^{(n)}, \\
\hat{\mathbf{x}}_i^{(n)} &= \sigma(\overrightarrow{\mathbf{x}}_i^{(n)} + \overleftarrow{\mathbf{x}}_i^{(n)} + \tilde{\mathbf{x}}_i^{(n)}),
\end{aligned} \tag{7}$$

where $\hat{\mathbf{x}}_i^{(0)} = \mathbf{x}_i^m$, $\overrightarrow{\mathbf{W}}^{(n)}, \overleftarrow{\mathbf{W}}^{(n)}, \widetilde{\mathbf{W}}^{(n)} \in \mathbb{R}^{D_e \times (D_x + D_h)}$ $\{\mathbf{b}_j^{(n)}\}_{j=1}^{N_e}, \widetilde{\mathbf{b}}^{(n)} \in \mathbb{R}^{D_e \times 1}$ are learnable parameters, and D_e is hyper-parameter. $\overrightarrow{\mathbf{x}}_i^{(n)}$ and $\overleftarrow{\mathbf{x}}_i^{(n)}$ are encoded features for out- and in- relationships respectively. $\tilde{\mathbf{x}}_i^{(n)}$ is the updated feature for itself. The final encoded feature $\hat{\mathbf{x}}_i^{(n)}$ is the sum of the above three features and σ is the activation function. By performing the gated graph convolution operation multiple iterations (N), semantic contexts representing multi-order relationships among vertices can be computed. Such semantic contexts are denoted as $\mathbf{X}^c = \{\mathbf{x}_i^c\}_{i=1}^K$.

Finally, for each vertex v_i, we concatenate its encoded spatial feature $\mathbf{p}_i$ mentioned before and its language-guided semantic context $\mathbf{x}_i^c$ to obtain the multimodal context $\mathbf{x}_i = [\mathbf{W}_p \mathbf{p}_i, \mathbf{x}_i^c]$, where $\mathbf{W}_p \in \mathbb{R}^{D_p \times 5}$.

3.3. Loss Function

The matching score between proposal o_i and expression L is defined as follows,

$$s_i = \text{L2Norm}(\mathbf{W}_{s0}\mathbf{x}_i) \odot \text{L2Norm}(\mathbf{W}_{s1}\mathbf{h}_g), \tag{8}$$

where $\mathbf{W}_{s0} \in \mathbb{R}^{D_s \times (D_p + D_x)}$ and $\mathbf{W}_{s0} \in \mathbb{R}^{D_s \times D_h}$ are transformation matrices.

Inspired by the deep metric learning algorithm for face recognition in [25], we adopt the triplet loss with online hard negative sample mining to train our CMRIN model. The triplet loss is defined as

$$loss = \max(s_{neg} + \Delta - s_{gt}, 0), \tag{9}$$

where s_{gt} and s_{neg} are the matching scores of the ground-truth proposal and the negative proposal respectively. The negative proposal is randomly chosen from the set of online hard negative proposals, $\{o_j | s_j + \Delta - s_{gt} > 0\}$, where Δ is the margin. During testing, we predict the target object by choosing the object proposal with the highest matching score.

4. Experiments

4.1. Datasets

We have evaluated our CMRIN on three commonly used benchmark datasets for grounding referring expressions (*i.e.*, RefCOCO [31], RefCOCO+ [31] and RefCOCOg [19]).

In RefCOCO, there are 50,000 target objects, collected from 19,994 images in MSCOCO [14], and 142,210 referring expressions. RefCOCO is split into train, validation, test A, and test B, which has 120,624, 10,834, 5,657 and 5,095 expression-target pairs, respectively. Test A includes images with multiple people, and test B includes images with multiple objects of other categories.

RefCOCO+ has 49,856 target objects collected from 19,992 images in MSCOCO, and 141,564 expressions collected from an interactive game interface. Different from RefCOCO, RefCOCO+ forbids absolute location descriptions in the expressions. It is split into train, validation, test A, and test B, which has 120,191, 10,758, 5,726 and 4,889 expression-target pairs, respectively.

RefCOCOg includes 49,822 target objects from 25,799 images in MSCOCO, and 95,010 long referring expressions collected in a non-interactive setting. RefCOCOg [22] has 80,512, 4,896 and 9,602 expression-target pairs for training, validation, and testing, respectively.

4.2. Evaluation and Implementation

The Precision@1 metric (the fraction of correct predictions) is used for performance evaluation. A prediction is considered to be a true positive if the top predicted proposal is the ground-truth one w.r.t the referring expression.

For a given dataset, we count the number of occurrences of each word in the training set. If a word appears more than five times, we add it to the vocabulary. Each word in the expression is initially an one-hot vector, which is further converted into a 512-dimensional embedding. Annotated regions of object instances are provided in RefCOCO, RefCOCO+ and RefCOCOg. The target objects in the three datasets belong to the 80 object categories in MSCOCO, but the referring expressions may mention objects beyond the 80 categories. In order to make the scope of target objects consistent with referring expressions, it is necessary to recognize objects in expressions, even when they are not within the 80 categories.

Inspired by the Bottom-Up Attention Model in [1] for image caption and visual question answering, we train ResNet-101 based Faster R-CNN [5, 23] over selected 1,460 object categories in the Visual Genome dataset [12],

			RefCOCO			RefCOCO+			RefCOCOg	
		feature	val	testA	testB	val	testA	testB	val	test
1	MMI [19]	vgg16	-	63.15	64.21	-	48.73	42.13	-	-
2	Neg Bag [22]	vgg16	76.90	75.60	78.00	-	-	-	-	68.40
3	CG [18]	vgg16	-	74.04	73.43	-	60.26	55.03	-	-
4	Attr [15]	vgg16	-	78.85	78.07	-	61.47	57.22	-	-
5	CMN [7]	vgg16	-	75.94	79.57	-	59.29	59.34	-	-
6	Speaker [31]	vgg16	76.18	74.39	77.30	58.94	61.29	56.24	-	-
7	Listener [32]	vgg16	77.48	76.58	78.94	60.50	61.39	58.11	69.93	69.03
8	Speaker+Listener+Reinforcer [32]	vgg16	79.56	78.95	80.22	62.26	64.60	59.62	71.65	71.92
9	VariContext [36]	vgg16	-	78.98	82.39	-	62.56	62.90	-	-
10	AccumulateAttn [4]	vgg16	81.27	81.17	80.01	65.56	68.76	60.63	-	-
11	ParallelAttn [37]	vgg16	81.67	80.81	81.32	64.18	66.31	61.46	-	-
12	MAttNet [30]	vgg16	80.94	79.99	82.30	63.07	65.04	61.77	73.04	72.79
13	Ours CMRIN	vgg16	84.02	84.51	82.59	71.46	75.38	64.74	76.16	76.25
14	MAttNet [30]	resnet101	85.65	85.26	84.57	71.01	75.13	66.17	78.10	78.12
15	Ours CMRIN	resnet101	**86.99**	**87.63**	**84.73**	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Table 1. Comparison with state-of-the-art approaches on RefCOCO, RefCOCO+ and RefCOCOg. The two best performing methods using VGG-16 are marked in red and blue.

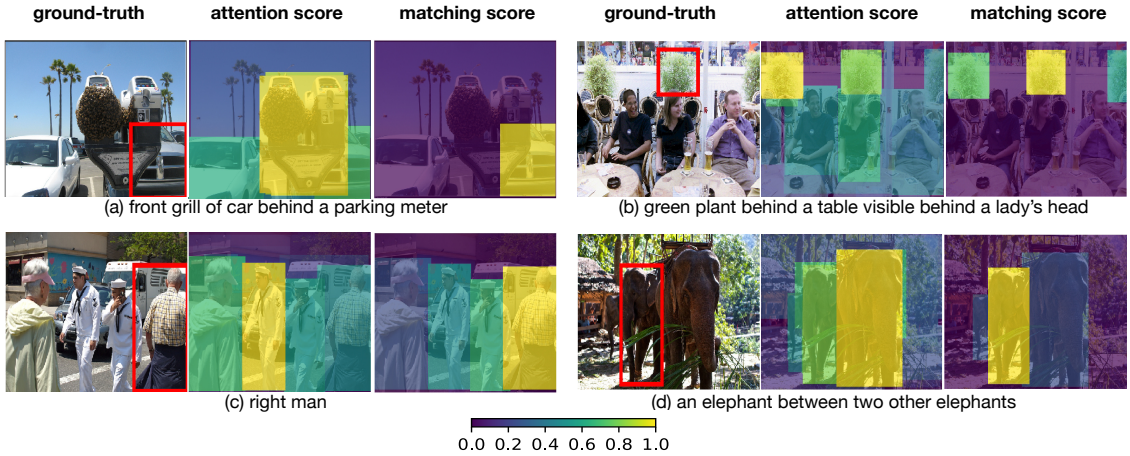

Figure 4. Qualitative results showing initial attention score (gate) maps and final matching score maps. We compute the score of a pixel as the highest score value among all proposals covering it, and normalize the score maps to 0 to 1. Warmer color indicates higher score.

excluding the images in the training, validation and testing sets of RefCOCO, RefCOCO+ and RefCOCOg. We combine the detected objects and the ground-truth objects provided by MSCOCO to form the final set of objects in the images. We extract the visual features of objects as the 2,048-dimensional output from the pool5 layer of the ResNet-101 based Faster R-CNN model. Since some previous methods use VGG-16 as the feature extractor, we also extract the 4,096-dimensional output from the fc7 layer of VGG-16 for fair comparison. We set the mini-batch size to 64. The Adam optimizer [10] is adopted to update network parameters with the learning rate set to 0.0005 initially. Margin is set to 0.1 in all our experiments.

4.3. Comparison with the State of the Art

We compare the performance of our proposed CMRIN against the state-of-the-art methods, including MMI [19], Neg Bag [22], CG [18], Attr [15], CMN [7], Speaker [31], Listener [32], VariContext [36], AccumulateAttn [4], ParallelAttn [37] and MAttNet [30].

4.3.1 Quantitative Evaluation

Table 1 shows quantitative evaluation results on RefCOCO, RefCOCO+ and RefCOCOg datasets. Our proposed CMRIN consistently outperforms existing methods across all the datasets by a large margin. Specially, CMRIN improves the average Precision@1 over validation and testing sets achieved by existing best-performing algorithm by 2.44%, 5.54% and 3.29% respectively on the RefCOCO, RefCOCO+ and RefCOCOg datasets when VGG-16 is used as the backbone. Our CMRIN significantly improves on the person category (testA of RefCOCO and RefCOCO+), which indicates that casting appearance attributes (*e.g.*, shirt, glasses and shoes) of a person as external relationships between person and appearance attributes can effectively distinguish the target person from other persons. After we switch to the visual features extracted by ResNet-101 based Faster R-CNN, the Precision@1 of our CMRIN

		RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
1	global langcxt+vis instance	79.05	81.47	77.86	63.85	69.82	57.80	70.78	71.26
2	global langcxt+global viscxt(2)	82.61	83.22	82.36	67.75	73.21	63.06	74.29	75.23
3	weighted langcxt+guided viscxt(2)	85.29	86.09	84.12	73.70	79.60	67.52	78.47	79.39
4	weighted langcxt+guided viscxt(1)+fusion	85.80	86.09	83.98	73.95	78.43	67.21	79.37	78.90
5	weighted langcxt+guided viscxt(3)+fusion	86.55	87.50	84.53	75.29	80.46	68.79	80.11	80.45
6	weighted langcxt+guided viscxt(2)+fusion	**86.99**	**87.63**	**84.73**	**75.52**	**80.93**	**68.99**	**80.45**	**80.66**

Table 2. Ablation study on variances of our proposed CMRIN on RefCOCO, RefCOCO+ and RefCOCOg. The number following the "viscxt" refers to the number of gated graph convolutional layers used in the model.

is further improved by another ∼3.61%. It improves the average Precision@1 over validation and testing sets achieved by MAttNet [30] by 1.29%, 4.38% and 2.45% respectively on the three datasets. Note that our CMRIN only uses the 2048-dimensional features from pool5, but MattNet uses multi-scale feature maps generated from the last convolutional layers of both the third and fourth stages.

4.3.2 Qualitative Evaluation

Visualizations of some samples along with their attention scores and matching scores are shown in Figure 4. They are generated from our CMRIN using ResNet-101 based Faster R-CNN features. Without relationship modeling, our CMRIN can identify the proposals appearing in the given expression (second columns), and it achieves this goal on the basis of single objects (*e.g.* the parking meter in Figure 4(a) and the elephant in full view in Figure 4(d) have higher attention scores). After fusing information from different modes and propagating multimodal information in the structured relation graph, it is capable of learning semantic contexts and locating target proposals (third columns) even when the target objects do not attract the most attention at the beginning. It is worth noting that our CMRIN learns semantic relations ("behind") for pairs of proposals with different spatial relations ("bottom right" between "car" and "parking meter" in Figure 4(a); "top" between "green plant" and "lady's head" in Figure 4(b)), which indicates that CMRIN is able to infer semantic relationships from the initial spatial relationships. In addition, CMRIN learns the context for target "elephant" (Figure 4(d)) from "two other elephants" by considering the relations from multiple elephants together. Moreover, multi-order relationships are learned through propagation in CMRIN, *e.g.*, the relationships ("right" in Figure 4(c)) between object pairs are propagated gradually to the target proposal (most "right man").

4.4. Ablation Study

Our proposed CMRIN includes CMRE and GGCN modules. To demonstrate the effectiveness and necessity of each module and further compare each module against its variants, we have trained five additional models for the comparison with the ResNet-101 based Faster R-CNN features. The results are shown in Table 2. As a baseline (row 1), we use the concatenation of instance-level visual features of objects and the location features as the visual features, and use the last hidden state of the expression encoding LSTM as the language feature, and then compute a matching score between every visual feature and the language feature. In comparison, a simple variant (row 2) that relies on a global visual context, which is computed by applying graph convolutional operations to the spatial relation graph, already outperforms the baseline. This demonstrates the importance of visual contexts. Another variant (row 3) with visual contexts computed in the language-guided visual relationship graph outperforms the above two versions. It captures the contexts by considering cross-modal information. By fusing the context of every word into the language-guided visual relationship graph, semantic contexts can be captured by applying gated graph convolutional operations (row 6, the final version of CMRIN). Finally, we explore the number of gated graph convolutional layers used in CMRIN. The 1-layer CMRIN (row 4) performs worse than the 2-layer CMRIN because it only captures contexts with first-order relationships. The 3-layer CMRIN (row 5) does not further improve the performance. One possible reason is that third-order relationships merely occur in the expressions.

5. Conclusion

In this paper, we have proposed an end-to-end Cross-Modal Relationship Inference Network (CMRIN) to compute and represent multimodal contexts for the task of grounding referring expressions in images. It consists of a Cross-Modal Relationship Extractor (CMRE) and a Gated Graph Convolutional Network (GGCN). CMRE extracts all the required information adaptively for constructing a language-guided visual relation graph with cross-modal attention. GGCN fuses information from different modes and propagates the fused information in the language-guided relation graph to obtain semantic contexts. Experimental results on three commonly used benchmark datasets show that our proposed method outperforms all existing state-of-the-art methods.

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 visual question answering. In *CVPR*, 2018.

[2] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2874–2883, 2016.

[3] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308. IEEE, 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 (CVPR)*, pages 7746–7755, 2018.

[5] 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 (CVPR)*, pages 770–778, 2016.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4418–4427. IEEE, 2017.

[8] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016.

[9] 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.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

[11] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2017.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

[15] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6985–6994, 2018.

[17] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[18] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, 2017.

[19] 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 (CVPR)*, pages 11–20, 2016.

[20] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[21] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2014.

[22] 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.

[23] 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.

[24] 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.

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, 2015.

[26] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[27] Abhinav Shrivastava and Abhinav Gupta. Contextual priming and feedback for faster r-cnn. In *European Conference on Computer Vision*, pages 330–348. Springer, 2016.

[28] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6857–6866, 2018.

[29] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. 2018.

[30] 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.

[31] 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.

[32] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speakerlistener-reinforcer model for referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, 2017.

[33] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018.

[34] Xingyu Zeng, Wanli Ouyang, Bin Yang, Junjie Yan, and Xiaogang Wang. Gated bi-directional cnn for object detection. In *European Conference on Computer Vision*, pages 354–369. Springer, 2016.

[35] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[36] 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 (CVPR)*, pages 4158–4166, 2018.

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