Modular Graph Attention Network for Complex Visual Relational Reasoning

Yihan Zheng[1,2], Zhiquan Wen[1], Mingkui Tan[1], Runhao Zeng[1,2], Qi Chen[1], Yaowei Wang[2(✉)], and Qi Wu[3]

[1] South China University of Technology, Guangzhou, China
yihanzheng7@gmail.com, sewenzhiquan@mail.scut.edu.cn, mingkuitan@scut.edu.cn
[2] PengCheng Laboratory, Shenzhen, China
wangyw@pcl.ac.cn
[3] The University of Adelaide, Adelaide, Australia
qi.wu01@adelaide.edu.au

Abstract. Visual Relational Reasoning is crucial for many vision-and-language based tasks, such as Visual Question Answering and Vision Language Navigation. In this paper, we consider reasoning on complex referring expression comprehension (c-REF) task that seeks to localise the target objects in an image guided by complex queries. Such queries often contain complex logic and thus impose two key challenges for reasoning: (i) It can be very difficult to comprehend the query since it often refers to multiple objects and describes complex relationships among them. (ii) It is non-trivial to reason among multiple objects guided by the query and localise the target correctly. To address these challenges, we propose a novel Modular Graph Attention Network (MGA-Net). Specifically, to comprehend the long queries, we devise a language attention network to decompose them into four types: basic attributes, absolute location, visual relationship and relative locations, which mimics the human language understanding mechanism. Moreover, to capture the complex logic in a query, we construct a relational graph to represent the visual objects and their relationships, and propose a multi-step reasoning method to progressively understand the complex logic. Extensive experiments on CLEVR-Ref+, GQA and CLEVR-CoGenT datasets demonstrate the superior reasoning performance of our MGA-Net.

1 Introduction

Visual relational reasoning often requires a machine to reason about visual and textual information and the relationships among objects before making a decision. This problem is crucial for many vision-and-language based tasks, such as visual question answering (VQA) [1–3] and vision language navigation (VLN)

Y. Zheng and Z. Wen—Contributed equally.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-69544-6_9) contains supplementary material, which is available to authorized users.

H. Ishikawa et al. (Eds.): ACCV 2020, LNCS 12627, pp. 137–153, 2021.
https://doi.org/10.1007/978-3-030-69544-6_9

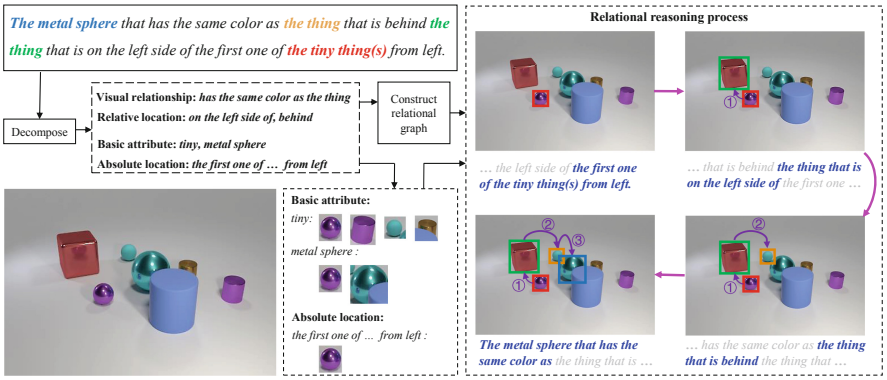

Fig. 1. An example of long-chain visual relational reasoning on CLEVR-Ref+ dataset [7]. The aim is to localise the target "metal sphere" based on the given complex query. To solve this, we propose to model the multi-step relationships via a relational graph step by step, and reason to the target.

[4–6]. However, reasoning can be very difficult because the visual and textual contents are often very complex. How to build a model to perform complex visual relational reasoning and how to validate the reasoning ability of such a model are still unclear.

Fortunately, we find the **complex** referring expression comprehension (c-REF) task [7,8] is a good test bed for visual reasoning methods. Specifically, c-REF requires a machine to reason over multiple objects and localise the target object in the image according to a complex natural language query (see Fig. 1). More critically, the complex visual and textual contents in this task can be a simulation of the complex real-world scenarios.

This task, however, is very challenging due to the following reasons: **a)** The query typically contains multiple types of information, such as the basic attributes, absolute location, visual relationship and relative location (see Fig. 1). It is non-trivial to understand the contents comprehensively in the complex query. **b)** Compared with the general referring expression comprehension (g-REF) task such as ReferCOCO [9,10], the query in c-REF often contains lots of visual relationships for multiple objects. It is very difficult to reason among multiple objects and localise the target correctly.

Recently, Liu *et al.* [7] found that state-of-the-art g-REF models like [11,12] failed to show promising reasoning performance on the c-REF task (*e.g.,* a new CLEVR-Ref+ dataset [7]), where the reasoning chain was long and complex. We find some methods [11,13] perform single-step reasoning to model the relationships between objects. However, the queries in real-world applications often contain complex logic, making them hard to be understood in such a one-step manner. For example, in Fig. 1, relational reasoning is a multi-step process: **Step 1**, select "*the first one of the tiny thing(s) from left*"; **Step 2**, given the object selected in Step 1, find "*the thing that is on the left side of*" them; **Step 3**,

find "*the thing that is behind*" the object selected in Step 2; **Step 4**, find "*the metal thing that has the same colour as*" the object selected in Step 3. Although some methods [14,15] attempt to update the object features for more than one step, it performs reasoning with the whole sentence without distinguishing different types of information. In this sense, it cannot handle the complex relationships in the query very well, thus leading to inferior results.

In this paper, we propose a Modular Graph Attention Network (MGA-Net), which considers different information in the query and models object relationships for multi-step visual relational reasoning. First, in order to comprehend the long queries, we propose a **language attention network** to decompose the query into four types, including the basic attributes, the absolute location, the visual relationship and the relative location. Second, based on the language representation of the basic attributes and absolute location, we propose an **object attention network** to find the object that is more relevant to the query. Third, to capture the complex logic in a query, we propose a **relational inference network**. In particular, we build a relational graph to represent the relationships between objects. Based on the graph, we propose a multi-step reasoning method based on Gated Graph Neural Networks (GGNNs) [16] to progressively understand the complex logic and localise the target object. We conduct experiments on CLEVR-Ref+ [7], GQA [8] and CLEVR-CoGenT [7] datasets, which contain multiple types of information in the queries and require relational reasoning to localise the target of interest.

Our main contributions are summarised as follows:

- To comprehend the complex natural language query, we decompose the query into four types and design a functional module for each type of information.
- We construct a relational graph among objects and propose a multi-step reasoning method based on Gated Graph Neural Networks (GGNNs) [16] to progressively understand the complex logic in the query. In this way, our method is able to effectively localise the target object especially when long-chain reasoning is required.
- MGA-Net achieves the best performance on three complex relational reasoning datasets, demonstrating the superiority of our proposed method.

2 Related Work

Visual Relational Reasoning. Many vision-and-language tasks require visual reasoning to focus on the referred object of the query, such as visual question answering (VQA) [3,17,18], visual language navigation (VLN) [4,19,20] and referring expression comprehension (REF) [11,13]. To well complete these high-level tasks, the model requires the ability of complex relational reasoning. Rather than treating the query as a single unit, recent works [21–23] decomposed the query into components and performed the reasoning with each component. Some works [11,24] exploited Neural Module Networks (NMNs) [25] to deal with different types of information. With clearly decomposing the query, the corresponding module networks are appropriately designed to achieve better reasoning ability.

Referring Expression Comprehension (REF). The REF task is to localise the referent in an image with the guidance of a given referring expression. For REF, several datasets such as RefCOCO [9], RefCOCO+ [9] and RefCOCOg [10] were released for research. However, as discussed in [15], the queries in these datasets did not require resolving relations. Moreover, recent research [26] argued that RefCOCO datasets were biased, which meant that we could obtain high accuracy without the queries.[1] To faithfully evaluate the reasoning ability of the models, Liu *et al.* [7] released CLEVR-Ref+ dataset which was approximately unbiased. In this paper, we focus on the complex referring expression comprehension (c-REF) and evaluate our model on CLEVR-Ref+, GQA and CLEVR-CoGenT datasets, which all reduce the statistical biases within the datasets. Different from traditional datasets, the above datasets require long-chain reasoning ability for understanding complex queries.

Graph Neural Networks (GNNs). GNNs [27,28] combine graph and neural networks to enable communication between the linked nodes and build informative representations. Many variants of GNNs, such as Graph Convolution Networks (GCN) [29], Graph Attention Network (GAT) [30] and Gated Graph Neural Networks (GGNNs) [16], were applied to various tasks [15,31,32]. Some recent works [13–15,33–35] performed relational reasoning using graph networks. Li *et al.* [33] constructed three graphs to represent the relations and updated the node features in the graphs with single-step reasoning. Wang *et al.* proposed LGRANs [13], which applied a graph attention network to better aggregate the information from the neighbourhood to perform the reasoning process. However, these methods performed single-step reasoning only, while many queries requiring multi-step reasoning to solve. DGA [34] and CMRIN [35] considered the relation for each pair of objects with a small relative distance, but they both ignored the attribute relation (e.g., two objects have the same colour). Moreover, LCGN [15] took the reasoning process into account and built a graph for multi-step reasoning. However, it encoded the query in a holistic manner without distinguishing different types of information, which was difficult to comprehend the complex query. Different from the above methods, our method distinguishes location relation and attribute relation, which refines the relation representation and contributes to complex relationship modelling. To perform multi-step reasoning, we construct two relational graphs and update the graph representations for multiple times based on the GGNNs [16].

3 Proposed Method

Our aim is to build a model to perform the relational reasoning guided by a complex query and then localise the target in an image. We choose the complex referring expression comprehension (c-REF) task to evaluate our model since

[1] Although RefCOCO datasets are biased and do not belong to the c-REF task, we conduct experiments on them and put the results into the supplementary material.

both complex reasoning and localisation are required in this task. Formally, given a natural language query r and its corresponding image I with N objects $\mathcal{O} = \{o_i\}_{i=1}^{N}$, the goal of c-REF is to identify the target object $o^\star$ by reasoning over the objects $\mathcal{O}$ guided by the query.

Due to the complexity of the queries, how to distinguish different types of information and how to reason among multiple objects guided by the query are very challenging. To deal with these challenges, we decompose the query into different types of information and design a functional module for each type of information. To capture the complex logic in a query, we construct a relational graph to represent the objects and their relationships. Moreover, we propose a multi-step reasoning method based on the relational graph to progressively understand the complex logic and identify the target object.

Our Modular Graph Attention Network (MGA-Net) composes of three components as shown in Fig. 2. First, the **language attention network** decomposes the query r into four types: basic attributes, absolute location, visual relationship and relative location, and obtains the corresponding language representations $\mathbf{s}^{att}$, $\mathbf{s}^{loc}$, $\mathbf{s}^{rel\_vis}$ and $\mathbf{s}^{rel\_loc}$. Second, the **object attention network** represents all candidate objects $\mathcal{O}$ with their visual features and spatial features, and obtains the objects that are relevant to r under the guidance of $\mathbf{s}^{att}$ and $\mathbf{s}^{loc}$. Third, the **relational inference network** constructs a relational graph among objects, and then updates node representations step by step via Gated Graph Neural Networks (GGNNs) guided by $\mathbf{s}^{rel\_vis}$ and $\mathbf{s}^{rel\_loc}$, respectively. Last, we match the updated object representations with the corresponding language representations to obtain the prediction.

3.1 Language Attention Network

A query in c-REF often describes multiple objects with their relationships and contains four types of information, including (1) basic attributes, which contain object category name, size, colour and material; (2) absolute location, which describes the position of the object in the image; (3) visual relationship, which represents the relationships (the same attribute or the inter-action) between objects; (4) relative location, which describes the displacement between objects.

To comprehend the query, some methods [25] adopted off-the-shelf language parser [36] to parse the query. However, as mentioned in [11], the external parser could raise parsing error, which affected the performance of REF. Therefore, instead of relying on the off-the-shelf language parser, we adopt self-attention mechanism to parse the query automatically.

To distinguish different types of information, we design a functional module for each type of information. Specifically, we represent a query with L words $r = \{w_l\}_{l=1}^{L}$ using the word embeddings $\{\mathbf{e}_l\}_{l=1}^{L}$, which can be obtained by using a non-linear mapping function or pre-trained word embeddings, such as GloVe [37]. With the word embeddings being the input of a Bi-LSTM model [38], we obtain the hidden state representations $\mathbf{h} = \{\mathbf{h}_l\}_{l=1}^{L}$, which are the concatenation of the forward and backward hidden vectors of the words. To calculate the attention score of each word in each module, we apply a fully connected layer to the hidden

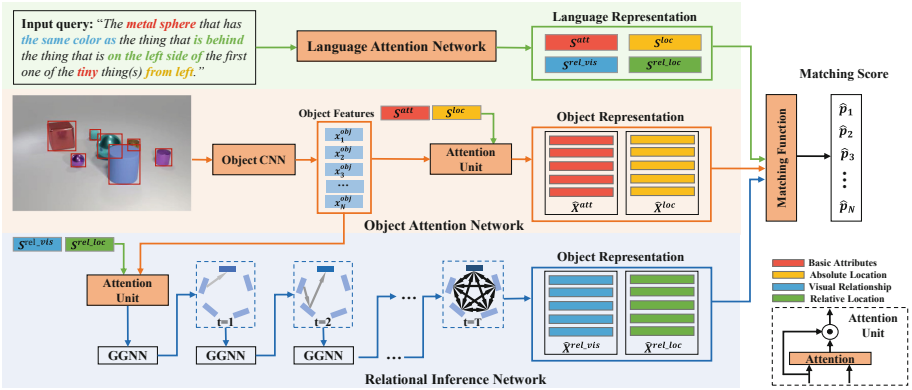

Fig. 2. Overview of Modular Graph Attention Network (MGA-Net) for visual relational reasoning. Our method contains three components. The language attention network decomposes the query into four types with an attention mechanism. The object attention network selects the related objects with their basic attributes and absolute locations. Based on the relational graph, the relational inference network models the complex relationships by a multi-step reasoning method. The final score is obtained by matching four object representations with their corresponding language representations.

state representations $\mathbf{h}$, and normalise the scores with a softmax function. In particular, for each word w_l, we calculate basic attributes attention a_l^{att}, absolute location attention a_l^{loc}, visual relationship attention $a_l^{rel\_vis}$ and relative location attention $a_l^{rel\_loc}$ as follows:

$$a_l^{type} = \frac{\exp\left(\mathbf{w}_a^{type\top}\mathbf{h}_l\right)}{\sum_{k=1}^{L}\exp\left(\mathbf{w}_a^{type\top}\mathbf{h}_k\right)}, \tag{1}$$

where $type \in \{att, loc, rel\_vis, rel\_loc\}$ and $\mathbf{w}_a^{type} \in \mathbb{R}^{d_w}$ denotes the parameters of each module and d_w is the dimension of the word embeddings. With the attention scores $\mathbf{a}^{att}$, $\mathbf{a}^{loc}$, $\mathbf{a}^{rel\_vis}$, $\mathbf{a}^{rel\_loc} \in \mathbb{R}^L$ at hand, we obtain the representation for each type of information as follows:

$$\mathbf{s}^{type} = \sum_{l=1}^{L} a_l^{type} \cdot \mathbf{e}_l. \tag{2}$$

With the help of the attention mechanism, we are able to learn the language representations *w.r.t.* the basic attributes, absolute location, relative location and visual relationship.

3.2 Object Attention Network

To localise the object with its properties (*i.e.,* the basic attributes and absolute location), we propose an object attention network. In particular, we represent

each object with its attribute feature and location feature. Then, we calculate attention scores for the objects with the guidance of the language representations $\mathbf{s}^{att}$ and $\mathbf{s}^{loc}$. Last, we update the object representations based on their basic attributes and absolute locations.

Basic Attributes Representation. The basic attributes module describes the object category, shape, colour, size and material of the object, which is relative to the visual features. The visual feature $\mathbf{u}_i$ for each object can be obtained by using some certain pre-trained feature extractor (*e.g.,* ResNet101 [39]). Then, we use a multi-layer perception (MLP) f^u with two hidden layers to obtain the basic attributes representation as $\mathbf{x}_i^{att} = f^u(\mathbf{u}_i)$.

Absolute Location Representation. The absolute location module describes the location information of the object. Supposing the width and height of the image are represented as $[W, H]$, and the top-left coordinate, bottom-right coordinate, width and height of object i are represented as $[x_{tl_i}, y_{tl_i}, x_{br_i}, y_{br_i}, w_i, h_i]$, then the spatial feature of object i is represented as a 5-dimensional vector $\mathbf{l}_i = \left[\frac{x_{tl_i}}{W}, \frac{y_{tl_i}}{H}, \frac{x_{br_i}}{W}, \frac{y_{br_i}}{H}, \frac{w_i \cdot h_i}{W \cdot H}\right]$. It denotes the top left and bottom right corner coordinates of the object region (normalised between 0 and 1) and its relative area (*i.e.*, the ratio of the bounding box area to the image area). Since the visual features may also indicate the object location from its background context, we also combine the visual features with the spatial feature, leading to the object location representation of object i as $\mathbf{x}_i^{loc} = \left[f^u(\mathbf{u}_i), f^l(\mathbf{l}_i)\right]$, where f^l is an MLP and $[\cdot, \cdot]$ denotes for concatenation.

Object Attention Module. Under the guidance of language representation $\mathbf{s}^{att}$ and $\mathbf{s}^{loc}$, the object attention module aims at finding the objects that are relevant to the given query. With the object representations $\mathbf{x}_i^{att}$ and $\mathbf{x}_i^{loc}$, the attention weights of object i is calculated as follows,

$$\begin{aligned} \tilde{a}_i^{o,obj} &= \mathbf{w}_o^{obj\top} \tanh\left(\mathbf{W}_{o,s}^{obj}\mathbf{s}^{obj} + \mathbf{W}_{o,x}^{obj}\mathbf{x}_i^{obj}\right), \\ a_i^{o,obj} &= \frac{\exp\left(\tilde{a}_i^{o,obj}\right)}{\sum_{j=1}^{N}\exp\left(\tilde{a}_j^{o,obj}\right)}, \end{aligned} \tag{3}$$

where $obj \in \{att, loc\}$. $\mathbf{W}_{o,s}^{obj} \in \mathbb{R}^{d_e \times d_w}$ and $\mathbf{W}_{o,x}^{obj} \in \mathbb{R}^{d_e \times d_o}$ are the parameters of two fully connected layers, which transform the language representations $\mathbf{s}^{obj}$ and the object representation $\mathbf{x}_i^{obj}$ into an embedding space, respectively. d_e is the dimension of the embedding space and d_o is the dimension of the object representation. $\mathbf{w}_o^{obj} \in \mathbb{R}^{d_e}$ is the parameters of the object attention module.

With the object attention weights $a_i^{o,att}$ and $a_i^{o,loc}$ at hand, we update the object representations by calculating

$$\hat{\mathbf{x}}_i^{obj} = a_i^{o,obj}\mathbf{x}_i^{obj}, obj \in \{att, loc\}. \tag{4}$$

In this way, the object representations are encoded with the language representations about basic attributes and the absolute location, respectively.

3.3 Relational Inference Network

To capture the complex logic in a query, we represent the input image as a graph, where nodes are objects and edges represent their relationships. Then, we adopt Gated Graph Neural Networks (GGNNs) to update the node representations by aggregating the information from neighbourhoods. However, for each node, a single-step updating cannot guarantee to capture the multi-step relationships between other nodes, making it hard to understand the complex logic in the query for reasoning. In this paper, we propose a multi-step updating method to progressively aggregate relational information for each node guided by the language representations $\mathbf{s}^{rel\_vis}$ and $\mathbf{s}^{rel\_loc}$. This updating method helps to understand the logic in the query and localise the object of interest correctly.

Graph Construction. We build a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ over the object set $\mathcal{O}$, where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the node set and $\mathcal{E} = \{e_{ij}\}_{i,j=1}^{N}$ is the edge set. Each node v_i corresponds to an object $o_i \in \{\mathcal{O}\}$ and each edge e_{ij} denotes the edge connecting objects o_i and o_j, which represents the relationships between the two objects.

Visual Relationship Representation. The visual relationship module describes the relationships between objects, referring to the attributes of two objects (*e.g.* A and B are in the same colour, size) or the interaction between objects (*e.g.* A holds B). To represent the visual relationship, we obtain the feature of object i by concatenating its visual feature and spatial feature $\mathbf{x}_i^{vis} = [\mathbf{u}_i, \mathbf{l}_i]$. Then, we use an MLP f^{rel_vis} to encode the features of two objects. The edge representation of the visual relationship between object i and object j is calculated as follows,

$$\mathbf{e}_{ij}^{rel\_vis} = f^{rel\_vis}\left(\left[\mathbf{x}_i^{vis}, \mathbf{x}_j^{vis}\right]\right). \tag{5}$$

Relative Location Representation. The relative location module describes the displacement between two objects which reflects the spatial correlation of objects. Here, we represent the spatial relation between two objects v_i and v_j as $\tilde{\boldsymbol{e}}_{ij} = [\frac{x_{tl_j} - x_{c_i}}{w_i}, \frac{y_{tl_j} - y_{c_i}}{h_i}, \frac{x_{br_j} - x_{c_i}}{w_i}, \frac{y_{br_j} - y_{c_i}}{h_i}, \frac{w_j \cdot h_j}{w_i \cdot h_i}]$, where $[x_{c_i}, y_{c_i}, w_i, h_i]$ is the centre coordinate, width and height of object i, and $\left[x_{tl_j}, y_{tl_j}, x_{br_j}, y_{br_j}, w_j, h_j\right]$ is the top-left coordinate, bottom-right coordinate, width and height of the j-th object, respectively. Considering the query like "A is to the left of B", the related object "B" plays an important role in location relationship understanding. We include the visual feature and spatial feature of object j for relative location representations. We obtain the edge representation of the relative location by calculating

$$\mathbf{e}_{ij}^{rel\_loc} = f^{rel\_loc}\left(\left[\tilde{\boldsymbol{e}}_{ij}, \mathbf{x}_j^{vis}\right]\right), \tag{6}$$

where f^{rel_loc} is an MLP. After learning the edge representation, we obtain the visual relational graph.

Multi-step Reasoning. Based on the constructed graph, we introduce Gated Graph Neural Networks (GGNNs) to iteratively update the node representations by aggregating the relational information and achieve the multi-step reasoning. GGNNs contain a propagation model learned with a gated recurrent update mechanism, which is similar to recurrent neural networks. For each object node in the graph, the propagation process of the t-th step is defined as follows:

$$\begin{aligned} \mathbf{z}_i^{rel,(t)} &= \tanh\left(\mathbf{a}_i^{rel\top}\left[\mathbf{h}_1^{rel,(t-1)};\ldots;\mathbf{h}_N^{rel,(t-1)}\right]\right), \\ \mathbf{h}_i^{rel,(t)} &= \text{GRUCell}\left(\mathbf{z}_i^{rel,(t)}, \mathbf{h}_i^{rel,(t-1)}\right), \end{aligned} \tag{7}$$

where $rel \in \{rel\_vis, rel\_loc\}$ and $\mathbf{a}_i^{rel}$ is the i-th row of a propagation matrix $\mathbf{A}^{rel} \in \mathbb{R}^{N\times N}$ that represents the propagation weights. GRUCell is the GRU update mechanism [40]. $\mathbf{h}_i^{rel,(t)}$ is the hidden state of the i-th object at step t.

At each time step, we update the node representations by aggregating the information from the neighbourhoods according to the propagation matrix $\mathbf{A}^{rel}$. However, updating for only one step fails to capture the multi-step relationships between other nodes. Thus, we propose a multi-step updating method, enabling each object to aggregate the relational information in the complex query and progressively understand the complex logic. After T time steps for propagation, the final representation for the i-th node can be obtained by:

$$\hat{\mathbf{x}}_i^{rel} = \mathbf{h}_i^{rel,(T)}. \tag{8}$$

Propagation Matrix. In Eq. (7), we need to compute the propagation matrix $\mathbf{A}^{rel}$ and the initial hidden state $\mathbf{h}_i^{rel,(0)}$ for each node. To this end, we devise an edge attention mechanism. Specifically, with the edge representation $\mathbf{e}_{ij}^{rel\_vis}$ and $\mathbf{e}_{ij}^{rel\_loc}$ in Eqs. (5) and (6), the edge attention for the visual relationship and relative location are calculated as follows,

$$\begin{aligned} \tilde{a}_{ij}^{rel} &= \mathbf{w}_e^{rel\top}\tanh\left(\mathbf{W}_{e,s}^{rel}\mathbf{s}^{rel} + \mathbf{W}_{e,x}^{rel}\mathbf{e}_{ij}^{rel}\right), \\ a_{ij}^{rel} &= \frac{\exp\left(\tilde{a}_{ij}^{rel}\right)}{\sum_{k\neq i}\exp\left(\tilde{a}_{ik}^{rel}\right)}, a_{ii}^{rel} = 0, \end{aligned} \tag{9}$$

where $rel \in \{rel\_vis, rel\_loc\}$. $\mathbf{W}_{e,s}^{rel} \in \mathbb{R}^{d_e\times d_w}$ and $\mathbf{W}_{e,x}^{rel} \in \mathbb{R}^{d_e\times d_r}$ are the parameters of two fully connected layers, which transform the expression $\mathbf{s}^{rel}$ and the edge representation $\mathbf{e}_{ij}^{rel}$ into an embedding space , respectively. d_e is the dimension of the embedding space and d_r is the dimension of the edge representation. $\mathbf{w}_e^{rel} \in \mathbb{R}^{d_e}$ are the parameters of the fully connected layers. With the edge attention mechanism, we obtain $a_{ij}^{rel\_vis}$ and $a_{ij}^{rel\_loc}$ to construct the propagation matrices. Then, the initial hidden state of the i-th object can be obtained by calculating $\mathbf{h}_i^{rel,(0)} = \sum_{j=1}^{N} a_{ij}^{rel}\mathbf{e}_{ij}^{rel}$, where $rel \in \{rel\_vis, rel\_loc\}$ and N is the number of nodes.

3.4 Matching Function and Loss Function

Matching Function. To find the target object, we need to compute a matching score for each object. Specifically, we devise a matching function to predict the final scores by matching the language representations and the corresponding object representations, which encode with the properties (in Eq. (4)) and relationships with other objects (in Eq. (8)). The matching score p_i^{type} between the language representation $\mathbf{s}^{type}$ and the object representation $\hat{\mathbf{x}}_i^{type}$ can be calculated as follows:

$$p_i^{type} = \tanh\left(\mathbf{W}_{m,s}^{type}\mathbf{s}^{type}\right)^{\top} \tanh\left(\mathbf{W}_{m,x}^{type}\hat{\mathbf{x}}_i^{type}\right), \tag{10}$$

where $type \in \{att, loc, rel\_vis, rel\_loc\}$, $\mathbf{W}_{m,s}^{type} \in \mathbb{R}^{d_e \times d_w}$ and $\mathbf{W}_{m,x}^{type} \in \mathbb{R}^{d_e \times d_e}$ are learnable parameters. Similar to the previous studies [11,13], we calculate four weights $[w^{att}, w^{loc}, w^{rel\_vis}, w^{rel\_loc}]$ to represent the contributions of different modules. We apply a fully connected layer to the vector $\mathbf{e} = \sum_{l=1}^{L} \mathbf{e}_l$. The calculation of the weights are as follows,

$$\left[w^{att}, w^{loc}, w^{rel\_vis}, w^{rel\_loc}\right] = \text{softmax}\left(\mathbf{W}_s\mathbf{e}\right), \tag{11}$$

where $\mathbf{W}_s \in \mathbb{R}^{4 \times d_w}$ is the parameter of the fully connected layer, and d_w is the dimension of the word embedding. For object i, the final matching score p_i is calculated by weighted summing up of the p_i^{type} with the four weights:

$$p_i = \sum_{type} w^{type} p_i^{type}. \tag{12}$$

Loss Function. To localise the referent among all objects in the image, we regard it as a multi-class classification task. The probability for object i being the referent is calculated as $\tilde{p}_i = \frac{\exp(p_i)}{\sum_{j=1}^{N}\exp(p_j)}$, where N is the number of object candidates in the image. We choose the cross-entropy loss as the loss function:

$$L = -\sum_{i=1}^{N} y_i \cdot \log\left(\tilde{p}_i\right), \tag{13}$$

where y_i is 1 when object i is the ground truth referent and 0 otherwise. We use the Adam [41] method to minimise the loss.[2]

4 Experiments

In this section, we evaluate the proposed method on a complex referring expression comprehension dataset (*i.e.,* CLEVR-Ref+ [7]). To evaluate the generalisation ability of our method, we further conduct experiments on CLEVR-CoGenT [7]. We also evaluate our method on a question answering dataset (*i.e.,* GQA [8]). Last, we perform ablation studies and visualisation analysis to verify the contributions of each module in our method.[3]

[2] We put the training algorithm into the supplementary material.

[3] We put the Implementation Details into the supplementary material.

4.1 Datasets

CLEVR-Ref+ [7] is a synthetic dataset whose images and queries are generated automatically. This dataset is approximately unbiased by employing a uniform sampling strategy. Moreover, it provides complex expressions that require strong visual reasoning ability to be comprehended.

GQA [8] is a real-world VQA dataset with compositional questions over images from Visual Genome data [42]. Like CLEVR-Ref+, GQA mitigates language priors and conditional biases for evaluating the visual reasoning capacity of models. Moreover, the questions in GQA include complex visual relationships among objects. In the GQA dataset, the grounding score is designed to check whether the model focuses on question and answer relevant regions within the images. Since MGA-Net focuses on visual reasoning (such as VQA and REF), the grounding scores of GQA is suitable for evaluating the performance of visual reasoning.

CLEVR-CoGenT [7] is a synthetic dataset, augmented from CLEVR [43] dataset. The queries are also complex, which require resolving relations. Moreover, it has two different conditions, such as Condition A and Condition B, which contain different object attributes.

4.2 Evaluation on CLEVR-Ref+

Comparison with State-of-the-Arts. We compare our MGA-Net with several state-of-the-art methods, including Stack-NMN [44], SLR [12], MAttNet [11], GroundeR [45] and LCGN [15]. From Table 1, our method outperforms all baselines. Specifically, MAttNet decomposes the queries into three parts (visual subject, location and relationship), but it ignores the issue of long-chain reasoning. Thus, MAttNet only achieves the accuracy of 60.9%. Beneficial from the multi-step reasoning based on the graph, LCGN leads to 14% improvement by using the graph network to get the context-aware representation. However, LCGN ignores to distinguish different types of information in the queries and encodes them in a holistic manner. Different from them, our MGA-Net considers the different information in the query and performs multi-step reasoning on the relational graph via GGNNs. With the detected bounding boxes as input, our proposed method achieves the accuracy of 80.1%. Moreover, using ground truth bounding boxes further improves the accuracy to 80.8%.

Effectiveness of Four Modules. MGA-Net decomposes queries into four parts, such as the basic attributes (att), absolute location (loc), relative location (rel_loc) and visual relationship (rel_vis). To evaluate the effect of each module, we conduct the ablation studies on CLEVR-Ref+ dataset. We use the ground truth bounding boxes as input and set the updating step of GGNNs to 3.

Quantitative Results. From Table 2, when only using basic attributes to localise the object (Row 1), the model achieves the accuracy of 62.10%. Row 2 shows the benefits brought by the absolute location module. By combining the visual relationship (Row 3) or the relative location module (Row 4), the accuracy improves

Table 1. Comparisons with state-of-the-arts on CLEVR-Ref+ in Accuracy.

Method	Accuracy (%)
Stack-NMN [44]	56.5
SLR [12]	57.7
MAttNet [11]	60.9
GroundeR [45]	61.7
LCGN [15]	74.8
MGA-Net (with detected bbox)	80.1
MGA-Net (with ground truth bbox)	**80.8**

Table 2. Impact of the four modules on CLEVR-Ref+.

Module	Accuracy (%)
att	62.10
att + loc	65.83
att + loc + rel_vis	72.81
att + loc + rel_loc	76.86
att + loc + rel_vis + rel_loc	**80.87**

Table 3. Comparisons of different models on GQA in terms of Grounding scores.

Method	Grounding score (%)
MAttNet [11]	56.73
LGRANs [13]	84.73
MGA-Net	**87.03**

Table 4. Comparisons with baselines on CLEVR-CoGenT (valA & B) in Accuracy.

Method	valA	valB
SLR [12]	0.63	0.59
MAttNet [11]	0.64	0.63
MGA-Net (with detected bbox)	0.82	0.76
MGA-Net (with ground truth bbox)	**0.83**	**0.78**

significantly, showing the benefit of the relational reasoning modules. Moreover, our MGA-Net with four modules achieves the best performance. These results demonstrate that distinguishing different types of information in the query is important for relational reasoning.

Visualisation. We visualise the four modules in the query in Fig. 3. We highlight the words which are correctly indicated by the corresponding modules according to the weight of each word. These results demonstrate that our proposed four modules have the ability to capture the corresponding phrases.

4.3 Evaluation on GQA

We evaluate MAttNet [11], LGRANs [13] and our MGA-Net on GQA. During training, we regard the mentioned object in the answer as the ground truth and train the models with the balanced training questions. In inference, all balanced validation questions are fed into the model to calculate the grounding scores. Such a score evaluates whether the model focuses on the regions of the image that are relevant to the questions and answers. Since the ground truth bounding boxes of the mentioned objects in the question and answer are not provided on the test-dev and test sets, we evaluate the methods on the validation set.

From Table 3, our MGA-Net outperforms the baselines. For LGRANs [13], it only introduces single-step reasoning, which is unsuitable for long-chain reasoning. Moreover, LGRANs considers the location relation only, and ignores the

Table 5. Impact of the updating step T. We report the accuracy (%) of our method with different values of T on CLEVR-Ref+ and CLEVR-CoGenT (valA & valB). We provide the grounding score (%) of our method with different values of T on GQA.

Dataset	CLEVR-Ref+		CLEVR-CoGenT (valA)		CLEVR-CoGenT (valB)		GQA
Setting	detected bbox	gt bbox	detected bbox	gt bbox	detected bbox	gt bbox	–
$T = 0$	75.81	76.51	76.00	76.24	71.37	72.32	86.01
$T = 1$	79.52	80.25	79.01	79.15	74.26	74.53	86.89
$T = 3$	**80.18**	**80.87**	**82.02**	**82.90**	**76.60**	**78.15**	87.03
$T = 5$	79.05	79.65	79.95	80.36	74.69	76.00	**87.93**

visual relations in the query. In contrast to the LGRANs, we first construct relational graphs among the objects regarding to the location and visual relations. Based on the relational graphs, we are able to conduct multi-step reasoning via GGNNs guided by language representations and thus achieves higher performance than LGRANs. Note that the excellent performance of MAttNet [11] in previous study relies on the attribute features and the phrase-guided "in-box" attention. However, these features cannot be provided by GQA. Besides, MAttNet only decomposes expressions into three modular components that are without multi-step reasoning, which makes it hard to perform the complex compositional reasoning. Thus, the performance of MAttNet degrades significantly on GQA.

4.4 Evaluation on CLEVR-CoGenT

To evaluate the generalisation ability, we train the model on the training set of Condition A and evaluate it on the validation set of Condition A and Condition B (*i.e.,* valA and valB). From Table 4, our MGA-Net outperforms SLR [12] and MAttNet [11] by a large margin. Specifically, in the "detection" setting, our method achieves the accuracy of 0.82 on valA and the accuracy of 0.76 on valB. When using ground truth bounding boxes, the performance of our method is further improved (0.83 on valA and 0.78 on valB). These results further demonstrate the superior generalisation ability of our proposed method.

Note that on the CLEVR-Ref+ and CLEVR-CoGenT datasets, our method achieves comparable performance between using ground truth bounding box and detected bounding box. The reason is that the scene of the image in these datasets is simple, and the objects in the scene are able to be detected accurately and easily.

4.5 Effectiveness of Multi-step Reasoning

Quantitative Results. To evaluate the multi-step reasoning, we verify it on our MGA-Net by setting different updating steps T in GGNNs. From Table 5, on all datasets, the models with GGNNs ($T > 1$) outperform the model without GGNNs ($T = 0$) significantly, which demonstrates the necessity and superiority of the relational reasoning.

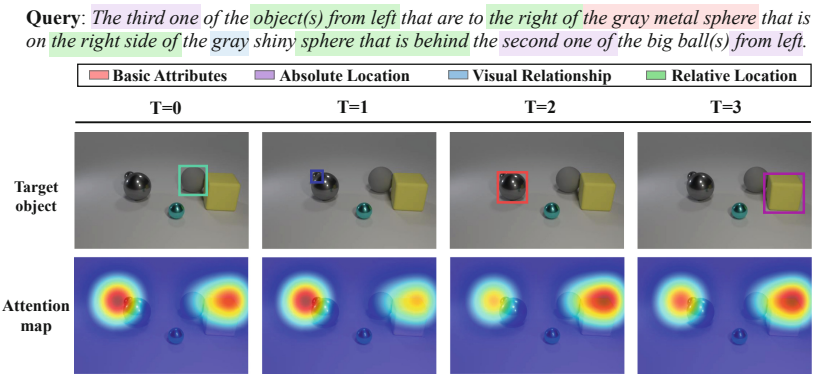

Fig. 3. An example of 3-step reasoning on complex referring expression comprehension. We visualise the attention maps of GGNNs and mark the target object by a bounding box for each step. With the guidance of the expression, the attentive image region changes over updating and the highlighted object corresponds to the ground truth.

Moreover, with the increasing of the updating steps (from $T = 1$ to $T = 3$), the performance of the MGA-Net further improves on all the datasets, which demonstrates the superiority of the multi-step reasoning. But MGA-Net with GGNNs $T = 3$ performs better than that with GGNNs $T = 5$ on CLEVR-Ref+ and CLEVR-CoGenT datasets. This implies that an appropriate number of updating steps helps to obtain the best performance on MGA-Net.

Visualisation. To further illustrate the effectiveness of our method on dealing with long-chain reasoning, we visualise the intermediate results of the model. We set the updating step $T = 3$ in GGNNs, and train the model with the ground truth bounding boxes as inputs on CLEVR-Ref+. Then, we obtain the initial nodes representations ($T = 0$) and the updated node representations in different updating steps ($T = 1, 2, 3$). To predict the score for each node, we match the node representations and the language representations. As shown in the visualisation results in Fig. 3, the attentive image region changes over updating and the highlighted object corresponds to the ground truth.

5 Conclusion

In this paper, we have proposed a new Modular Graph Attention Network (MGA-Net) for complex visual relational reasoning. To cover and represent the textual information, we decompose the complex query into four types and design a module for each type. Meanwhile, we construct the relational graphs among objects. Based on the relational graphs, we devise a graph inference network with GGNNs to update the graph representations step by step. Our method

encodes the multi-step relationships among objects and then reasons to the target. Promising results demonstrate the effectiveness and the superior visual relational reasoning ability of our method.

Acknowledgement. This work was partially supported by the Key-Area Research and Development Program of Guangdong Province 2019B010155002, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT-07X183, Fundamental Research Funds for the Central Universities D2191240.

References

1. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (AAAI) (2018)
2. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. In: International Conference on Learning Representations (ICLR) (2018)
3. Mao, J., Gan, C., Kohli, P., Tenenbaum, J.B., Wu, J.: The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision. In: International Conference on Learning Representations (ICLR) (2019)
4. Anderson, P., et al.: Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3674–3683 (2018)
5. Wang, X., et al.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6629–6638 (2019)
6. Nguyen, K., Dey, D., Brockett, C., Dolan, B.: Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12527–12537 (2019)
7. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: CLEVR-REF+: diagnosing visual reasoning with referring expressions. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4185–4194 (2019)
8. Hudson, D.A., Manning, C.D.: GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6700–6709 (2019)
9. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: ReferitGame: referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 787–798 (2014)
10. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11–20 (2016)
11. Yu, L., et al.: MAttNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1307–1315 (2018)
12. Yu, L., Tan, H., Bansal, M., Berg, T.L.: A joint speaker-listener-reinforcer model for referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3521–3529 (2017)

13. Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., Hengel, A.v.d.: Neighbourhood watch: referring expression comprehension via language-guided graph attention networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1960–1968 (2019)
14. Bajaj, M., Wang, L., Sigal, L.: G3raphGround: graph-based language grounding. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4281–4290 (2019)
15. Hu, R., Rohrbach, A., Darrell, T., Saenko, K.: Language-conditioned graph networks for relational reasoning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10294–10303 (2019)
16. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.S.: Gated graph sequence neural networks. In: International Conference on Learning Representations (ICLR) (2016)
17. Norcliffe-Brown, W., Vafeias, S., Parisot, S.: Learning conditioned graph structures for interpretable visual question answering. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 8334–8343 (2018)
18. Chang, S., Yang, J., Park, S., Kwak, N.: Broadcasting convolutional network for visual relational reasoning. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 754–769 (2018)
19. Huang, H., Jain, V., Mehta, H., Ku, A., Magalhaes, G., Baldridge, J., Ie, E.: Transferable representation learning in vision-and-language navigation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 7404–7413 (2019)
20. Ke, L., et al.: Tactical rewind: self-correction via backtracking in vision-and-language navigation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6741–6749 (2019)
21. Johnson, J., et al.: Inferring and executing programs for visual reasoning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2989–2998 (2017)
22. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: end-to-end module networks for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 804–813 (2017)
23. Cao, Q., Liang, X., Li, B., Li, G., Lin, L.: Visual question reasoning on general dependency tree. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7249–7257 (2018)
24. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1115–1124 (2017)
25. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 39–48 (2016)
26. Cirik, V., Morency, L., Berg-Kirkpatrick, T.: Visual referring expression recognition: what do systems actually learn? In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 781–787 (2018)
27. Gori, M., Monfardini, G., Scarselli, F.: A new model for learning in graph domains. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), vol. 2, pp. 729–734. IEEE (2005)
28. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Networks **20**, 61–80 (2008)

29. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017)
30. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
31. Zeng, R., et al.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 7094–7103 (2019)
32. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. In: AAAI Conference on Artificial Intelligence (AAAI), pp. 11021–11028 (2020)
33. Li, L., Gan, Z., Cheng, Y., Liu, J.: Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10313–10322 (2019)
34. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4644–4653 (2019)
35. Yang, S., Li, G., Yu, Y.: Cross-modal relationship inference for grounding referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4145–4154 (2019)
36. Zhu, M., Zhang, Y., Chen, W., Zhang, M., Zhu, J.: Fast and accurate shift-reduce constituent parsing. In: ACL, vol. 1, pp. 434–443 (2013)
37. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
38. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
40. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734 (2014)
41. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
42. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**, 32–73 (2017)
43. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1988–1997 (2017)
44. Hu, R., Andreas, J., Darrell, T., Saenko, K.: Explainable neural computation via stack neural module networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 55–71 (2018)
45. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 817–834 (2016)