Bottom-Up Shift and Reasoning for Referring Image Segmentation

Sibei Yang[1*†] Meng Xia[2*] Guanbin Li[2] Hong-Yu Zhou[3] Yizhou Yu[3,4†]
[1]ShanghaiTech University [2]Sun Yat-sen University
[3]The University of Hong Kong [4]Deepwise AI Lab

Abstract

Referring image segmentation aims to segment the referent that is the corresponding object or stuff referred by a natural language expression in an image. Its main challenge lies in how to effectively and efficiently differentiate between the referent and other objects of the same category as the referent. In this paper, we tackle the challenge by jointly performing compositional visual reasoning and accurate segmentation in a single stage via the proposed novel Bottom-Up Shift (BUS) and Bidirectional Attentive Refinement (BIAR) modules. Specifically, BUS progressively locates the referent along hierarchical reasoning steps implied by the expression. At each step, it locates the corresponding visual region by disambiguating between similar regions, where the disambiguation bases on the relationships between regions. By the explainable visual reasoning, BUS explicitly aligns linguistic components with visual regions so that it can identify all the mentioned entities in the expression. BIAR fuses multi-level features via a two-way attentive message passing, which captures the visual details relevant to the referent to refine segmentation results. Experimental results demonstrate that the proposed method consisting of BUS and BIAR modules, can not only consistently surpass all existing state-of-the-art algorithms across common benchmark datasets but also visualize interpretable reasoning steps for stepwise segmentation. Code is available at https://github.com/incredibleXM/BUSNet.

1. Introduction

The intersection of vision and language has attracted growing interests in academia, where many methods [1, 3, 25] have been proposed to promote a better understanding of these two modalities. Existing vision-and-language approaches can be roughly divided into two types based on their designing principles, *i.e.*, multimodal fusion and representation learning, and language-conditioned visual reasoning. In contrast to the former which focuses more on how to learn joint representations from multiple modalities, the latter reasoning based approaches usually are not only more effective in complex scenes but also can provide an explainable decision-making process.

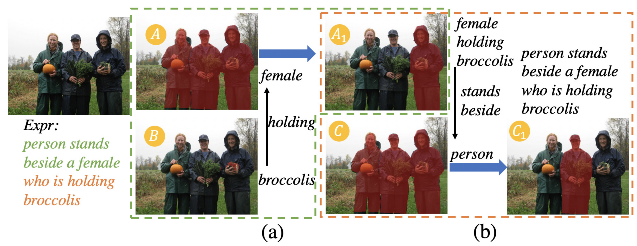

Figure 1. The Bottom-Up Shift (BUS) for referring image segmentation. BUS performs stepwise visual reasoning from the entity "*broccolis*" to "*female*" to "*person*". At each step, it first identifies the objects corresponding to the entity and then differentiates between the identified objects by the relational reasoning.

However, as one of the most fundamental vision-and-language tasks, Referring Image Segmentation (RIS) [8] has not been well addressed in previous research works from the second perspective (*i.e.*, reasoning). Existing visual reasoning based methods [19, 46] for RIS mainly resort to a two-stage pipeline, where they first detect and segment the object instances and then perform reasoning over feature vectors of both object instances and their relationships. However, the two-stage solution inevitably faces the problems of slow inference speed and has poor generalization [23]. What is worse, the relational and spatial priors in images that are essential for visual reasoning are lost when conducting reasoning over feature vectors of those object instances. On the other hand, most existing works [9, 10, 15] on RIS mainly focus on learning multimodal contextual representations in a single stage. Generally, one-stage RIS methods have fast inference speed but are inferior in handling complex visual scenes and expres-

*Equal contribution.
†Corresponding authors. This work was partially supported by National Key Research and Development Program of China (No.2020YFC2003902), National Natural Science Foundation of China (No.61976250 and No.U1811463), and the Guangdong Basic and Applied Basic Research Foundation (No.2020B1515020048).

sions because they lack sufficient visual reasoning capability [23]. For example (see Figure 1), without visual reasoning, the model can not distinguish the referred "*person*" from others in the image.

In this paper, we aim to empower the one-stage RIS with the ability to conduct visual reasoning and take advantages of both one-stage and two-stage methods. The two-stage methods rely on explicit object instances and their relationships to conduct visual reasoning; however, there is no explicit object-level information in one-stage RIS. Therefore, we propose that capturing visual scenes' constituents and their relationships is the key to perform visual reasoning in one-stage RIS. In Figure 1(a), given the linguistic structure ("*female*"-"*holding*"-"*broccolis*") of the referring expression ("*A female is holding broccolis*"), we can first align visual regions A and B with the nouns "*female*" and "*broccolis*" respectively, and then shift region A to A_1 by considering its relationship "*holding*" to region B. By the process, the referred "*female*" is located with interpretable reasoning steps. Moreover, we can perform bottom-up shift and reasoning to identify the referent hierarchically for complex expressions. As shown in Figure 1(b), we further segment the referred "*person*" by the following two steps. First, we find region C with respect to the noun "*person*". Then, we shift region C to region C_1 by considering its relationships "*stands beside*" to the identified region A_1. Also, we can refine the visual region B by considering its inverse relationship "*be held*" with A_1. In addition to finding the referent, bidirectional shifts for a pair of relationship and inverse relationship help to segment other mentioned objects.

To realize the above concepts and operations, we propose a Bottom-Up Shift (BUS) module to introduce visual reasoning to one-stage RIS. Specifically, BUS first parses the expression as a language graph and then analyzes hierarchical reasoning steps from the graph. In the language graph, each node and directed edge represent a specific noun phrase and the type of semantic relationship from the object node to the subject node, respectively. Then, BUS conducts bottom-up visual reasoning on the entire image following the reasoning steps. Particularly, we decompose the compositional visual reasoning process into pairwise relational shifts on edges and integration on nodes. The pairwise relational shift performs visual reasoning for a single edge by passing messages between its two nodes according to the type of this edge, where relationship-based convolutional operations implement the message passing.

Moreover, how to accurately segment the referent from a coarse localization also plays a vital role in RIS. Previous works [9, 10, 15] usually incorporate multi-level features to refine the details of segmentation results. However, these approaches either neglect the low-level visual details or capture incomplete interactions between multiple levels via a one-way fusion. In this paper, we propose a Bidirectional Attentive Refinement (BIAR) module to integrate low-level visual features and high-level semantic ones. Specifically, the top-down branch is responsible for capturing semantic-related visual details, while the bottom-up pathway helps to equip multi-level semantic features with the captured details. However, directly incorporating the low-level visual features into high-level semantic ones may bring irrelevant noise, because low-level visual features contain visual details of the entire image. Thus, we propose an attention mechanism to incorporate the details relevant to the referent selectively.

In summary, this paper has following contributions:

- A Bottom-Up Shift (BUS) module is proposed to empower one-stage referring image segmentation with the ability to perform explainable visual reasoning. The BUS can not only distinguish the referent from other objects of the same category as the referent but also segment other mentioned entities in the expression.
- A Bidirectional Attentive Refinement (BIAR) module is proposed to segment the referent from a coarse localization accurately. BIAR integrates low-level visual features and high-level semantic ones via a two-way attentive message passing, which improves the segmentation accuracy.
- BUS and BIAR are integrated into a Bottom-Up Shift and Reasoning Network (BUSNet). Experimental results demonstrate that BUSNet not only outperforms existing state-of-the-art methods and achieves significant performance gains over referring expression reasoning models, but also generates interpretable visualizations for stepwise reasoning and segmentation.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation aims to segment all pixels of objects from predefined categories. Fully convolutional network (FCN) [22] and its variants have become dominant in semantic segmentation. To alleviate the down-sampling issue, DeepLab [5] replaces the traditional convolutions with atrous convolutions in FCNs to enlarge the receptive filed of convolutions without losing spatial details. Different approaches have been introduced to aggregate multi-scale context. For example, DeepLabv2 [6] and PSPNet [48] capture objects and context at multiple scales via pyramid atrous convolutions and pyramid spatial pooling, respectively. Besides, low-level visual features have been integrated to complement the detailed information [16, 29].

2.2. Referring Image Comprehension and Segmentation

Referring image comprehension aims to locate a bounding box that corresponds to the object referred by an ex-

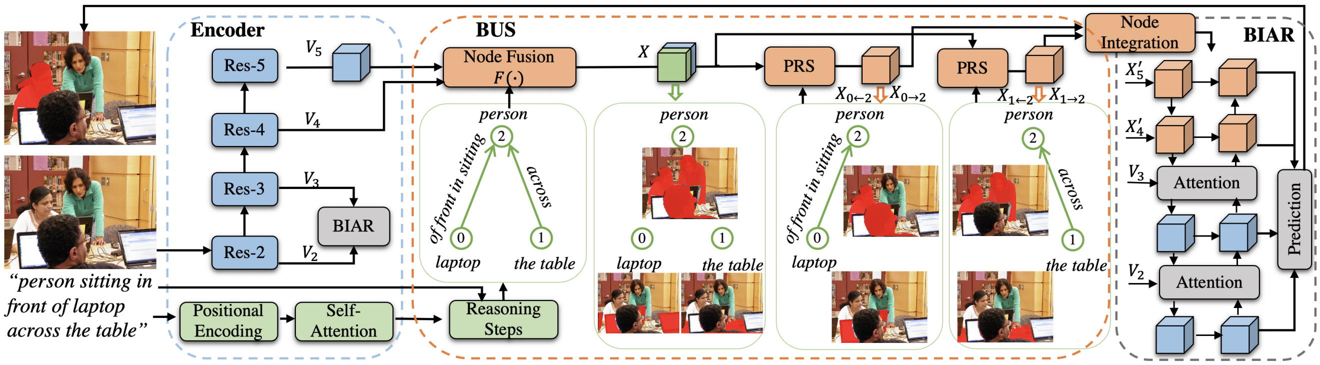

Figure 2. An overview of our Bottom-Up Shift and Reasoning Network (BUSNet). Encoder extracts multi-level visual features $\{\boldsymbol{V}_i\}_{i=2}^{5}$ and language features from the input image and expression. Bottom-Up Shift (BUS) module performs explainable visual reasoning on the high-level visual features $\{\boldsymbol{V}_i\}_{i=4}^{5}$ via the Pairwise Relational Shift (PRS) and integration operations, and the outputs $\{\boldsymbol{X}'_i\}_{i=4}^{5}$ of it embed the relevant information of the referent. Bidirectional Attentive Refinement (BIAR) module integrates the low-level visual features $\{\boldsymbol{V}_i\}_{i=2}^{3}$ and the high-level semantic ones $\{\boldsymbol{X}'_i\}_{i=4}^{5}$ to refine the segmentation results.

pression. Appearance information, spatial locations and attributes of objects as well as the relationships between objects are jointly utilized to help distinguish the referent from other objects [38, 41, 42, 46]. Different from referring image comprehension, referring image segmentation aims to locate the referent with a precise mask instead of a bounding box. Some approaches [19, 46] attempt to predict the masks of the referents by directly utilizing the referring image comprehension models. However, these methods often have slow inference speed and poor generalization ability [31]. Mainstream approaches address referring image segmentation in a more straightforward one-stage architecture, where they encode multimodal representations and then predict pixel-wise segmentation mask in a fully convolutional manner [8]. The multimodal LSTM [18], dynamic filter [26], recurrent refinement [15] and text-guided exchange [10] are proposed to achieve a better fusion for the multi-level visual features and sequential textual representations. Recently, some approaches resort to attention mechanisms to enhance the key information [33] or capture dependencies between these two modalities [9, 23, 44].

2.3. Explainable Visual Reasoning on Relationships

Visual reasoning is developed to perform multi-step inferences on complex visual content in a visual scene, and the inferences are over the scene's constituents and their relationships. Relation network [32] captures pairwise relationships between every pair of visual regions to perform relational reasoning. Some works [11, 37, 43] resort to attention mechanisms to perform multi-step reasoning. Neural module networks [2, 13, 27, 7] decompose compositional reasoning into a sequence of sub-tasks and address these sub-tasks in independent modules. Neural-symbolic approaches [45, 24] first extract symbolic representations, based on which the symbolic programs are then executed.

Visual reasoning has also been exploited for relational modelling in recent advances in referring image comprehension and segmentation. DGA [39] performs relational reasoning by dynamically identifying a sequence of compound objects. NMTree [19] and SGMN [40] perform tree- or graph-structured referring expression reasoning via neural modules. However, their reasoning methods are based on explicit object instances which are not available for one-stage referring image segmentation. CGAN [23] and LSPN [41] are proposed to perform stepwise reasoning over the entire image to recognize instance-level semantic differences. However, their grouped attention reasoning and relational propagation are implicit and too coarse compared to ours, which cannot provide a clear explanation for the reasoning.

3. Bottom-Up Shift and Reasoning

The overall framework of the Bottom-Up Shift and Reasoning Network (BUSNet) is shown in Figure 2. Given an input image and an input expression, we first extract the visual feature maps at multiple levels and textual representations using the visual backbone and language encoder, respectively (in Section 3.1). For each high-level visual feature map, we then feed it together with the textual representations to the proposed Bottom-Up Shift module (BUS) to identify the referent. The BUS module performs stepwise reasoning via the pairwise relational shift and integration (in Section 3.2). Next, to refine the segmentation results, the bidirectional attentive refinement is proposed to integrate multi-level features by passing attentive messages in top-down and bottom-up pathways (in Section 3.3).

3.1. Image and Language Encoders

Image Encoder. Following prior works [9, 44], we adopt DeepLab ResNet101 as the visual backbone, and extract features of $\{Res\text{-}2, Res\text{-}3, Res\text{-}4, Res\text{-}5\}$ from the input image I as the visual feature maps $\{\boldsymbol{V}_2, \boldsymbol{V}_3, \boldsymbol{V}_4, \boldsymbol{V}_5\}$, where $\boldsymbol{V}_i$ corresponds to the feature of $Res\text{-}i$ with $i \in \{2, 3, 4, 5\}$. Besides, referring expressions often describe absolute locations of referents, such as "right pizza" and "the elephant in the middle". Therefore, we also encode 8-dim spatial coordinates [8] of visual feature maps as representations for the image. For each visual feature map $\boldsymbol{V}_i$, we denote its corresponding spatial feature map as $\boldsymbol{P}_i$.

Language Encoder. Given the expression $L = \{l_t\}_{t=1}^{T}$, we first extract the GloVe [30] word embedding $\boldsymbol{w}_t$ of each word l_t. Similar to [44], we use the assemble of individual word vectors instead of the entire sentence vector to represent the whole expression. To make use of the order of the sequence, we encode the relative positions of words in the expression using the positional encoding [34]. For each word l_t, we sum up its positional embedding $\boldsymbol{pos}_t$ and word embedding $\boldsymbol{w}_t$ to obtain a position-aware vector which is denoted as $\boldsymbol{w}'_t \in \mathbb{R}^{D_w \times 1}$. To further enhance the language representations, we capture dependencies between the words via the self-attention mechanism [35], and the new word representation $\boldsymbol{h}_t \in \mathbb{R}^{D_h \times 1}$ of word l_t is computed as follows:

$$\begin{aligned} \boldsymbol{h}_t &= \sum_{i=1}^{T} \alpha_{t,i} \boldsymbol{v}_i, \\ s.t.\ \boldsymbol{\alpha}_t &= Softmax([\boldsymbol{q}_t^T \boldsymbol{k}_i]_{i=1}^{T}), \end{aligned} \tag{1}$$

where $\boldsymbol{q}_t = \boldsymbol{W}_q \boldsymbol{w}'_t$, $\boldsymbol{k}_i = \boldsymbol{W}_k \boldsymbol{w}'_i$, $\boldsymbol{v}_i = \boldsymbol{W}_v \boldsymbol{w}'_i$. $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v \in \mathbb{R}^{D_h \times D_w}$ are linear transformation matrices. $\alpha_{t,i}$ denotes the ith element of the attention vector $\boldsymbol{\alpha}_t$.

Consider that each high-level visual feature map is fed to the bottom-up shift module (Section 3.2) respectively for stepwise reasoning, we ignore the index subscript of $\boldsymbol{V}$ and $\boldsymbol{P}$ for simplicity of demonstration.

3.2. Bottom-Up Shift

The Bottom-Up Shift (BUS) module achieves explainable visual reasoning in one-stage referring image segmentation by performing stepwise reasoning on the entire visual feature map. In practice, BUS aligns visual constituents (*i.e.*, visual regions and their relationships) with linguistic components explicitly following hierarchical reasoning steps. Specifically, we first represent the reasoning steps as a hierarchical order of traversal on a language graph which is parsed from the expression. Then, we perform stepwise inferences on the graph's edges and nodes via the pairwise relational shift and integration modules.

3.2.1 Analysis of Reasoning Steps

The reasoning steps to locate the referent are indicated by the referring expression which describes how objects modify and interact with the referent. Inspired by [20, 40, 41], we first represent the expression as a language graph which is a directed acyclic graph with a referent node whose outdegree is zero. A node and a directed edge of the graph respectively correspond to a noun phrase and the linguistic relationship (*e.g.*, a preposition/verb phrase) from object to subject. Then, we collect these linguistic relationships and define a set of types of linguistic relationships, such as "ride" and"sit". Next, we convert the linguistic relationships of the edges to different types. Formally, the final language graph $\mathcal{G}$ of the expression L is defined as $\mathcal{G} = (\mathcal{O}, \mathcal{E})$, where $\mathcal{O} = \{o_n\}_{n=1}^{N}$ and $\mathcal{E} = \{e_k\}_{k=1}^{K}$ are sets of nodes and directed edges respectively. Specifically, each node o_n is associated with a entity (*i.e.* noun/noun phrase), and the referent node is denoted as o_{ref}. Each directed edge $e_k = (e_k^{(s)} \in \mathcal{O}, e_k^{(r)}, e_k^{(o)} \in \mathcal{O})$ from $e_k^{(o)}$ to $e_k^{(s)}$ can be regarded as a triplet containing the subject node $e_k^{(s)}$, the type of relationship $e_k^{(r)}$, and the object node $e_k^{(o)}$. And we denote the set of edges whose subject node is o_n as $\mathcal{E}_n$.

Thanks to the graph-structure representation of the expression, we can simplify the compositional reasoning into a multi-step inference on nodes and edges of the graph. We define the reasoning steps by running a reverse breadth-first traversal on the graph from its referent node and adopting the traversed order as the reasoning order of the nodes. The traversed order essentially guarantees that when we get the node for reasoning, all the other nodes that modify that node already have been processed. The order of multi-step reasoning over nodes is from the bottom to the top. The hierarchical reasoning of the example in Figure 2 is from "*laptop*" and "*the table*" to "*person*".

3.2.2 Stepwise Inference

We perform stepwise inference following the extracted reasoning steps (*i.e.*, the traversed order on the language graph). Each node of the language graph corresponds to a visual region in the image, and the stepwise inference is proposed to identify the correct visual region of each node by conducting relational reasoning over edges.

First, we obtain nodes' initial feature maps which encode nodes' initial spatial locations in the image. The initial feature maps can be obtained by fusing the visual feature map $\boldsymbol{V} \in \mathbb{R}^{H \times W \times D_v}$, the spatial feature map $\boldsymbol{P} \in \mathbb{R}^{H \times W \times 8}$ and the language representations of nodes. Specifically, we extract the language representation of node o_n as the mean of the word embeddings of this node's noun phrase. For each node o_n with the language representation $\bar{\boldsymbol{h}}_n$, its multimodal feature map $\boldsymbol{X}_n \in \mathbb{R}^{H \times W \times D_x}$ can be computed as follows:

$$\boldsymbol{X}_n = Conv_v([\boldsymbol{V}; \boldsymbol{P}]) * Tile(\boldsymbol{W}_{\bar{h}} \bar{\boldsymbol{h}}_n) \tag{2}$$

where $*$ is the element-wise multiplication, $[;]$ is a concatenation operation, $Conv_x$ and $\boldsymbol{W}_{\bar{h}} \in \mathbb{R}^{D_x \times D_h}$ are the convolutional layer and learnable matrix with $tanh$ as the activation function, respectively. $Tile$ means to tile vectors to produce a feature map with size of $H \times W \times D_x$. The above fusion process can be simplified into $\boldsymbol{X}_n = F(\boldsymbol{V}, \boldsymbol{P}, o_n)$, where $F(\cdot)$ stands for all fusion operations.

Next, we shift nodes' initial spatial locations in the image to the correct ones by performing stepwise reasoning over the relationships between nodes, *i.e*., edges. We process nodes step by step following the traversed order. Similarly, we suppose node o_n is processed as a subject node in the current step. o_n is modified by the nodes that connects to it, *i.e*., the object nodes of edges $\mathcal{E}_n$ (see Section 3.2.1). We first individually perform relational reasoning over each edge in $\mathcal{E}_n$ via the Pairwise Relational Shift (PRS), and then integrate the results of node o_n from all connected edges $\mathcal{E}_n$ via an average pooling operation. For ease of presentation, we first present the integration from edges here and introduce more details about the PRS module later in Section 3.2.3. For the node o_n with initial feature map $\boldsymbol{X}_n$ and connected edges $\mathcal{E}_n$, its updated feature map $\boldsymbol{X}'_n$ is computed as follows:

$$\begin{aligned} \boldsymbol{X}_{n \leftarrow m}, \boldsymbol{X}_{n \rightarrow m} &= PRS^{(3)}(\boldsymbol{X}_n, e_k^{(r)}, \boldsymbol{X}'_m), \\ \boldsymbol{X}'_n &= \frac{\sum_{o_m \in e_k^{(o)} \& e_k \in \mathcal{E}_n} \boldsymbol{X}_{n \leftarrow m} + \boldsymbol{X}_n}{|\mathcal{E}_n| + 1} \end{aligned} \tag{3}$$

where PRS denotes the PRS module and $PRS^{(3)}$ means that the PRS module is applied three times iteratively, $e_k \in \mathcal{E}_n$ represents the directed edge whose subject node is o_n, $o_m \in e_k^{(o)}$ is the object node of edge e_k, $\boldsymbol{X}'_m$ is the updated feature map at node o_m, and $|\mathcal{E}_n|$ is the number of edges in $\mathcal{E}_n$. Note that the traversed order has guaranteed that the features of the node o_m already have been updated to $\boldsymbol{X}'_m$ when we start to process node o_n. Also, we can further update the $\boldsymbol{X}'_m$ using $\boldsymbol{X}_{n \rightarrow m}$ to refine the information at node o_m. Accordingly, the updated feature map $\boldsymbol{X}'_n$ will be used to update for upper nodes.

By performing the reasoning from bottom to up, we can finally obtain the updated feature map $\boldsymbol{X}'_{ref}$ of the uppermost node (*i.e*., the referent node o_{ref}), which encodes all the relational information from its child nodes. The reasoning process can be explicitly explained by the hierarchical inference order and the decoded attention maps of feature maps at nodes (see Section 4.4).

3.2.3 Pairwise Relational Shift

Pairwise Relational Shift (PRS) performs relational reasoning over a single edge by passing messages between two nodes according to the type of the linguistic relationship of this edge. The message from one node can help the other to refine its corresponding visual region or distinguish the region from other similar regions. Inspired by the predicate operator [14], we implement the message passing by designing a group of relationship-based convolution operations. We learn the weights of convolution kernels respectively for each type of linguistic relationship because the relational shifts of the same relationship often remain similar between varying nodes. For instance, given the relationship "below", we should focus the attention below the object when locating the subject. Accordingly, we should move our attention above the object when the relationship "ride" is given.

The inputs to PRS module include the type of the edge and the feature maps of both the subject node and the object node. PRS then outputs the updated representations of these two nodes by incorporating the influence of the type of edge connecting them. Given a single edge $e = (e^{(s)}, e^{(r)}, e^{(o)})$ and the feature maps $\boldsymbol{X}_s$ and $\boldsymbol{X}_o$ of the subject and the object nodes, the new feature maps $\boldsymbol{X}_{s \leftarrow o}$ and $\boldsymbol{X}_{s \rightarrow o}$ are computed as follows:

$$\begin{aligned} \boldsymbol{A}_{s \leftarrow o} = \gamma(Conv_r^{-1}(\boldsymbol{X}_o)), \boldsymbol{X}_{s \leftarrow o} = F(\boldsymbol{A}_{s \leftarrow o} \odot \boldsymbol{V}, \boldsymbol{P}, e^{(s)}), \\ \boldsymbol{A}_{s \rightarrow o} = \gamma(Conv_r(\boldsymbol{X}_s)), \boldsymbol{X}_{s \rightarrow o} = F(\boldsymbol{A}_{s \rightarrow o} \odot \boldsymbol{V}, \boldsymbol{P}, e^{(o)}), \end{aligned} \tag{4}$$

where $Conv_r$ and $Conv_r^{-1}$ are stacked convolutional layers corresponding to the edge type $e^{(r)}$ and its inverse type, γ denotes the $tanh$ activation function, $\odot$ stands for the pixel-wise multiplication, $\boldsymbol{V}$ and $\boldsymbol{P}$ correspond to the visual feature map and spatial feature map (see in Section 3.1), and the $F(\cdot)$ is the fusion function (see Section 3.2.2). The attention map of node $e^{(s)}$ is $\boldsymbol{A}_{s \leftarrow o} \in \mathbb{R}^{H \times W}$, which is obtained from the object node's feature map $\boldsymbol{X}_o$, is used to fuse for the new feature map $\boldsymbol{X}_{s \leftarrow o}$ of node $e^{(s)}$. Note that we can iteratively apply the same PRS module multiple times to refine the attention maps progressively by replacing the inputs $\boldsymbol{X}_s$ and $\boldsymbol{X}_o$ with the new feature maps $\boldsymbol{X}_{s \leftarrow o}$ and $\boldsymbol{X}_{s \rightarrow o}$.

3.3. Bidirectional Attentive Refinement

Multi-level features have been integrated to improve the segmentation accuracy in previous works. These works [44, 9] first process visual feature maps at multiple levels respectively and repeatedly, and then integrate the results from different levels. However, the repeated treatment to multi-level feature maps severely increase the computational cost. More importantly, the characteristics of the visual feature maps at different levels are not fully utilized. High-level features reveal semantic content, while low-level feature provide structural details. Therefore, we apply the visual reasoning (*i.e*., BUS module) on the high-level visual feature maps $\{\boldsymbol{V}_4, \boldsymbol{V}_5\}$ to obtain the referent's semantic features $\{\boldsymbol{X}'_4, \boldsymbol{X}'_5\}$, and further aggregate the low-level visual feature maps $\{\boldsymbol{V}_2, \boldsymbol{V}_3\}$ with the high-level ones to acquire more visual details.

We utilize both top-down and bottom-up pathways to refine the multi-level feature maps $\{\boldsymbol{V}_2, \boldsymbol{V}_3, \boldsymbol{V}_4, \boldsymbol{X}'_4, \boldsymbol{X}'_5\}$ progressively. The higher-level semantic features provide the semantic and spatial information of the referent to the lower-level visual features in the top-down pathway,

Method	Type	UNC val	UNC testA	UNC testB	UNC+ val	UNC+ testA	UNC+ testB	G-Ref val
Fusion and Refinement								
RMI [18]	one-stage	44.33	44.74	44.63	29.91	30.37	29.43	34.40
DMN [26]	one-stage	49.78	54.38	45.13	38.88	44.22	32.29	36.76
RRN+DCRF [15]	one-stage	55.33	57.26	53.95	39.75	42.15	36.11	36.45
CMSA+DCRF [44]	one-stage	58.32	60.61	55.09	43.76	47.60	37.89	39.98
STEP [4]	one-stage	60.04	63.46	57.97	48.19	52.33	40.41	46.40
CMPC+DCRF [10]	one-stage	61.36	64.53	59.64	49.56	53.44	43.23	49.05
BRINet+DCRF [9]	one-stage	61.35	63.37	59.57	48.57	52.87	42.12	48.04
LSCM+DCRF [12]	one-stage	61.47	64.99	59.55	49.34	53.12	43.50	48.05
Explainable Reasoning								
MAttNet [8]	two-stage	56.51	62.37	51.70	46.67	52.39	40.08	-
NMTree [18]	two-stage	56.59	63.02	52.06	47.40	53.01	41.56	-
CGAN [23]	one-stage	59.25	62.37	53.94	46.16	51.37	38.24	46.54
Ours BUSNet	one-stage	62.56	65.61	60.38	50.98	56.14	43.51	49.98
Ours BUSNet+DCRF	one-stage	**63.27**	**66.41**	**61.39**	**51.76**	**56.87**	**44.13**	**50.56**

Table 1. Comparison with state-of-the-art referring image segmentation methods on UNC, UNC+ and G-Ref datasets using overall IoU (%). DCRF denotes DenseCRF post-processing.

while the lower-level features with the details of the image are then integrated into the higher-level ones. For notation consistency, we denote the $\{\boldsymbol{V}_2, \boldsymbol{V}_3, \boldsymbol{V}_4, \boldsymbol{X}'_4, \boldsymbol{X}'_5\}$ as $\{\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \boldsymbol{G}_4, \boldsymbol{G}_5\}$. In the top-down branch, the features are computed as follows:

$$\boldsymbol{A}_i^{td} = \sigma(Conv_c(Conv_a(\boldsymbol{G}_i) + Conv_b(Up(\boldsymbol{G}_{i+1}^{td}))))$$
$$\boldsymbol{G}_i^{td} = \begin{cases} Conv_i(\boldsymbol{G}_i), & \text{if } i \in \{5\} \\ Conv_i(\boldsymbol{G}_i + Up(\boldsymbol{G}_{i+1}^{td})), & \text{if } i \in \{4\} \\ Conv_i(\boldsymbol{A}_i^{td} \odot \boldsymbol{G}_i + Up(\boldsymbol{G}_{i+1}^{td})), & \text{if } i \in \{1,2,3\} \end{cases} \tag{5}$$

where $\sigma(\cdot)$ represents the sigmoid function, $Conv$s are the convolutional operations for feature processing, Up is the upsampling operation and $\odot$ stands for the pixel-wise multiplication. Note that the low-level visual features contain details of the entire image and may bring irrelevant noise to the referent, thus, we compute an attention map $\boldsymbol{A}_i^{td} \in \mathbb{R}^{H_i \times W_i}$ to extract the attentive details of the referent. Then, the bottom-up passing is applied to the features $\{\boldsymbol{G}_1^{td}, \boldsymbol{G}_2^{td}, \boldsymbol{G}_3^{td}, \boldsymbol{G}_4^{td}, \boldsymbol{G}_5^{td}\}$ to obtain the bidirectional attentive features $\{\boldsymbol{G}'_1, \boldsymbol{G}'_2, \boldsymbol{G}'_3, \boldsymbol{G}'_4, \boldsymbol{G}'_5\}$. The bottom-up branch shares a similar computation process with the top-down one.

Finally, we upsample and sum up the bidirectional attentive feature maps to predict the segmentation mask [44].

4. Experiments

4.1. Experimental Setup

Datasets. To evaluate the proposed algorithm, we have conducted experiments on three common benchmark datasets, including UNC [47], UNC+ [47], and G-Ref [25]. Concretely, the UNC dataset has 142,209 expressions referring to 50,000 objects in 19,994 images. And the UNC+ dataset contains 19,992 images with 141,564 expressions for 49,856 objects. The absolute location descriptions are forbidden in UNC+. The G-Ref dataset collected from MSCOCO via the Amazon's Mechanical Turk, which consists of 104,560 expressions referring to 54,822 objects in 26,711 images.

Implementation Details. For a fair comparison with previous works [9, 23], we employ DeepLab ResNet-101 pretrained on Pascal VOC dataset as the visual backbone. Input images are resized to 320×320. For the language encoder, we use GloVe [30] pretrained on Common Crawl 840B tokens as our initial word embeddings and set the maximum length of the referring expression to 20. For the linguistic relationships, we collect 31, 30, and 33 types of relationships for UNC, UNC+ and G-Ref datasets, respectively. The dimensions of word representations and multi-level visual feature maps are set to 512 (*i.e.*, $D_h = D_w = 512$). Also, the dimensions of features in BUS module are set to 512. We train the network with RAdam optimizer [21]. The initial learning rate is $2.5e^{-4}$ and the weight decay is $5e^{-4}$. Weighted binary cross-entropy loss and Dice Loss [28] are applied over all pixels during training. DenseCRF is adopted to refine the segmentation masks following prior works [10, 12].

The overall intersection-over-union (IoU) and Prec@X metrics are used to evaluate the performance of referring image segmentation models [9, 10]. The overall IoU is the total intersection areas divided by the total union areas over all the test samples. The Prec@X is the percentage of prediction masks whose IoU score are higher than a given threshold X, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.2. Comparison with State-of-the-Arts

We compare the proposed model with state-of-the-art methods in referring image segmentation, and comparison results are shown in Table 1. Our model consistently outperforms all the state-of-the-art models (SOTAs) across all three benchmark datasets by large margins. Our model improves the average performance of overall IoU achieved by

	Method	$prec$@0.5	$prec$@0.6	$prec$@0.7	$prec$@0.8	$prec$@0.9	overall IoU
1	Baseline	39.09	32.22	26.10	15.54	3.20	35.25
2	+ Postional Encoding (with Self-attention)	44.65	38.42	31.92	18.85	5.97	38.92
3	+ Positional Encoding + GloVe = Multi-level	45.78	40.59	33.64	20.03	6.32	40.39
4	Multi-level + FPN	46.82	41.90	35.33	21.59	7.03	41.15
5	Multi-level + ConvLSTM	48.05	43.29	36.72	22.87	8.22	43.08
6	Multi-level + BIAR = Refinement	50.73	44.12	38.84	26.52	9.58	44.13
7	Refinement + BUS-1	54.93	48.72	42.07	29.92	10.60	46.81
8	Refinement + Concat-1	44.37	40.16	32.45	19.83	6.75	39.95
9	Refinement + BUS-1 w/o Type	51.13	44.35	38.28	24.39	8.84	43.86
10	Refinement + BUS-3	57.09	52.95	47.84	37.92	14.21	49.97
11	Refinement + BUS-4	55.94	51.13	46.77	36.87	13.52	48.58
12	Refinement + BUS-2	56.81	51.20	46.74	37.98	15.24	49.98

Table 2. Ablation study on the validation set of G-Ref using $prec@X$ (%) and overall IoU (%). All the models use the same visual backbone DeepLab ResNet-101, and no any post-processing is applied.

existing best-performing methods by 1.66%, 2.09%, and 1.51% on the UNC, UNC+ and G-Ref datasets, respectively.

Compared with SOTAs for explainable reasoning, the proposed method achieves significant performance gains on all the splits by 3.39%-7.45%, which demonstrates the effectiveness of our visual reasoning method in referring image segmentation. Recently, CGAN [23] is proposed for one-stage referring expression reasoning, which has the same setting and motivation as ours. Our model significantly surpasses CGAN by large margins of 5.17%, 5.66% and 4.02% on UNC, UNC+ and G-Ref respectively, which indicates that our model can better equip one-stage referring image segmentation with visual reasoning capability. Moreover, the proposed method also improves the overall IoU achieved by two-stage methods (*i.e*., MAttNet [46] and NMTree [19]) by 6.47% and 3.60% on UNC and UNC+ datasets, even when MAttNet and NMTree have more powerful pretrained backbones [12, 23]. Besides, the inference speed of the proposed BUSNet is about five times faster than that of the two-stage methods on the same hardware.

Compared with SOTAs from the multimodal fusion and progressive refinement perspective, our models improve the overall IoU consistently across all the benchmarks. Note that the fusion and refinement models usually have higher performance than the reasoning ones [40, 23]; however, they do not have the internal reasoning process.

4.3. Ablation Study

To evaluate the effectiveness of the language encoder, the proposed BIAR and BUS modules, we have trained 11 additional models for the comparison. The results are shown in Table 2.

Baseline and Language Encoder. The baseline model (row 1) simply fuses the visual feature maps, the spatial feature maps and the language representations of the expression at multiple levels, and predicts the segmentation mask from the fused features. The language representation is extracted from word embeddings of words in the expression via the mean-pooling operation, and the word embeddings are learned from scratch. As shown in row 2, the language encoder with positional encoding and self-attention improves the overall IoU of baseline by 3.67%, which demonstrates the effectiveness of the encoding method. Moreover, adopting the pretrained word embedding of GloVe (row 3) will further improve the overall IoU by 1.47%.

Multi-Level Refinement of BIAR. We conduct ablation study on multi-level refinement and evaluate models with different refinement methods. As shown in row 3 to row 6 of Table 2, the FPN [17] (row 4), ConvLSTM [36] (row 5) and our BIAR (row 6) have better performance than the multi-level baseline (row 3) that sums up the multi-level features as one, which indicates the effectiveness of progressive refinement for multi-level features. The bidirectional refinement manner of our BIAR encodes the attentive details of the image to the high-level semantic features, which outperforms one-way FPN and ConvLSTM by 2.98% and 1.05%, respectively.

Visual Reasoning of BUS. We further equip the model with the reasoning ability and examine different settings of BUS. The results are shown in row 6 to row 12 of Table 2. The BUS-1 (row 7) model applies the BUS module to a single visual feature map $\boldsymbol{V}_5$, which outperforms the reasoning baseline (row 6) without any inference module by 2.68% in terms of overall IoU. The performance gain clearly validates the effectiveness of the inference module for referring image segmentation. Concat-1 (row 8) and BUS-1 w/o Type (row 9) are two variants of the BUS module. Concat-1 neglects the non-local relationships between visual regions by replacing the pairwise relational shift (PRS) of BUS as a simple concatenation of the nodes' feature maps and edge's language features, while BUS w/o Type ignores the types of edges by learning shared convolutional parameters of PRS over all edge types. The worse performance of Concat-1 and BUS w/o Type demonstrates that the incorrect message passing over nodes adversely affect the model, and adopting the PRS module with different edge types is crucial for learning appropriate relational shifts. Finally, we explore the multi-level BUS reasoning (row 10 to row 12). The BUS-2, BUS-3 and BUS-4 models perform the BUS reasoning on the visual feature maps $\{\boldsymbol{V}_4, \boldsymbol{V}_5\}$, $\{\boldsymbol{V}_3, \boldsymbol{V}_4, \boldsymbol{V}_5\}$,

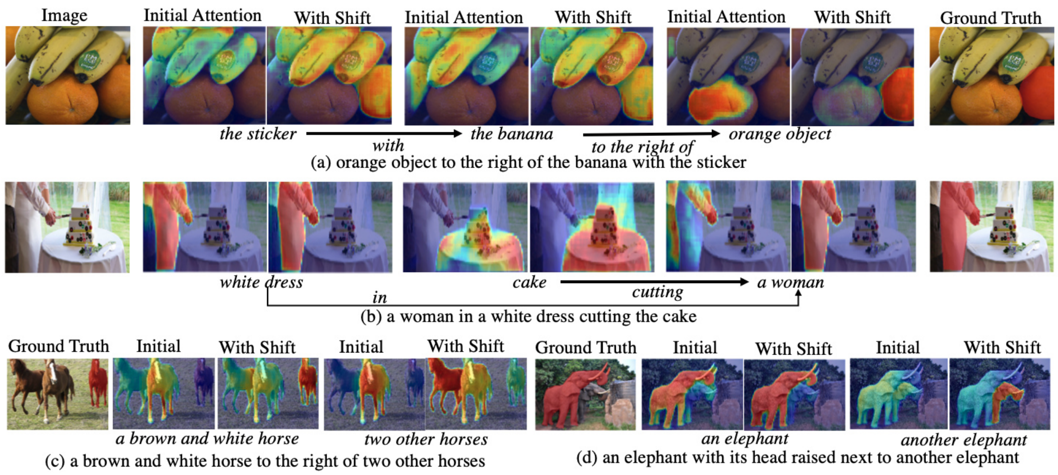

Figure 3. Qualitative results showing reasoning structures and attention maps of reasoning steps. Warmer color indicates higher score.

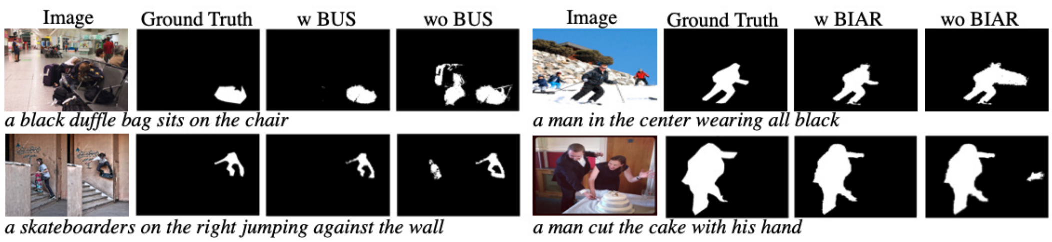

Figure 4. Qualitative results showing the effects of the BUS and BIAR modules.

and $\{\boldsymbol{V}_2, \boldsymbol{V}_3, \boldsymbol{V}_4, \boldsymbol{V}_5\}$, respectively. The BUS-2 and BUS-3 improves the overall IoU of BUS-1 (row 7) by 3.17% roughly, which reveals the importance of multi-level inference for objects of different scales. The BUS-4 does not further improve the performance because it loses details (embedded in visual features) of the image by performing the BUS on all the levels.

4.4. Qualitative Evaluation

We visualize reasoning processes and segmentation masks to explore in-depth insights into the proposed model. The reasoning structures and attention maps of the reasoning steps are shown in Figure 3. Specifically, we feed the multimodal feature maps $\{\boldsymbol{X}_n\}$ and updated feature maps $\{\boldsymbol{X}'_n\}$ into the decoder of segmentation to generate the initial attention maps and shifted attention maps, respectively. The qualitative evaluation results demonstrate that our model can generate interpretable intermediate processes for stepwise segmenting the referent. In Figure 3(a), our model performs bottom-up reasoning from the "*sticker*" to "*banana*" to "*orange object*". First, it shifts the initial attention of "*banana*" to the final one via the relational shift and successfully grounds "*the banana with the sticker*" in the image. Then, it identifies the target "*orange object*" that is to the right of the located "*banana*". In Figure 3(b), our model performs hierarchical inference from both "*white dress*" and "*cake*" to "*a woman*". As shown in the initial attention of "*a woman*", without visual reasoning, the model pays more attention to the arm of the man. By reasoning over relationships between "*a woman*" and other entities, the model finds the referred "*a woman*" who is in a "*white dress*" and cutting "*the cake*". In addition to locating the referent, the proposed model can also identify other entities mentioned in the expression. Two examples are shown In Figure 3(c) and (d). The model not only finds the referred "*a brown and white horse*" and "*an elephant*" but also other objects (*i.e.*, "*two other horses*" and "*another elephant*").

To demonstrate the effects of BUS and BIAR modules, we visualize the segmentation masks that are predicted with or without them, and the results are shown in Figure 4. Without the BUS module, the objects of the same category cannot be distinguished precisely. With the BIAR module, the boundaries and details of segmentation masks are closer to the one of ground truth.

5. Conclusion

In this paper, we propose Bottom-Up Shift (BUS) module to disambiguate between the referent and objects of the same category as the referent and introduce Bidirectional Attentive Refinement (BIAR) module to refine the coarse localization from visual details. The proposed method consisting of BUS and BIAR not only outperforms all SOTAs but also achieves significant gains over existing visual reasoning models in referring image segmentation.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2018. 1

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 39–48, 2016. 3

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Int. Conf. Comput. Vis.*, pages 2425–2433, 2015. 1

[4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Int. Conf. Comput. Vis.*, pages 7454–7463, 2019. 6

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*, 2015. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 2

[7] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *Eur. Conf. Comput. Vis.*, pages 53–69, 2018. 3

[8] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Eur. Conf. Comput. Vis.*, pages 108–124. Springer, 2016. 1, 3, 4, 6

[9] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4424–4433, 2020. 1, 2, 3, 5, 6

[10] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10488–10497, 2020. 1, 2, 3, 6

[11] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *Int. Conf. Learn. Represent.*, 2018. 3

[12] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *Eur. Conf. Comput. Vis.*, pages 59–75. Springer, 2020. 6, 7

[13] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Int. Conf. Comput. Vis.*, pages 2989–2998, 2017. 3

[14] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6867–6876, 2018. 5

[15] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5745–5753, 2018. 1, 2, 3, 6

[16] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. pages 1925–1934, 2017. 2

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017. 7

[18] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Int. Conf. Comput. Vis.*, pages 1271–1280, 2017. 3, 6

[19] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Int. Conf. Comput. Vis.*, pages 4673–4682, 2019. 1, 3, 7

[20] Daqing Liu, Hanwang Zhang, Zheng-Jun Zha, Meng Wang, and Qianru Sun. Joint visual grounding with language scene graphs. *arXiv preprint arXiv:1906.03561*, 2019. 4

[21] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Int. Conf. Learn. Represent.*, April 2020. 6

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015. 2

[23] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *ACM Int. Conf. Multimedia*, pages 1274–1282, 2020. 1, 2, 3, 6, 7

[24] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *Int. Conf. Learn. Represent.*, 2019. 3

[25] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 1, 6

[26] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Eur. Conf. Comput. Vis.*, pages 630–645, 2018. 3, 6

[27] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4942–4950, 2018. 3

[28] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571. IEEE, 2016. 6

[29] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. *Int. Conf. Comput. Vis.*, pages 1520–1528, 2015. 2

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. 4, 6

[31] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Int. Conf. Comput. Vis.*, pages 4694–4703, 2019. 3

[32] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Adv. Neural Inform. Process. Syst.*, pages 4967–4976, 2017. 3

[33] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Eur. Conf. Comput. Vis.*, pages 38–54, 2018. 3

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.* 4

[35] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018. 4

[36] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Adv. Neural Inform. Process. Syst.*, pages 802–810, 2015. 7

[37] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *Eur. Conf. Comput. Vis.*, pages 451–466. Springer, 2016. 3

[38] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4145–4154, 2019. 3

[39] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Int. Conf. Comput. Vis.*, pages 4644–4653, 2019. 3

[40] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9952–9961, 2020. 3, 4, 7

[41] Sibei Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *Eur. Conf. Comput. Vis.*, pages 589–605. Springer, 2020. 3, 4

[42] Sibei Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[43] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21–29, 2016. 3

[44] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10502–10511, 2019. 3, 4, 5, 6

[45] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Adv. Neural Inform. Process. Syst.*, pages 1031–1042, 2018. 3

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. pages 1307–1315, 2018. 1, 3, 7

[47] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85. Springer, 2016. 6

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2881–2890, 2017. 2