Bi-directional Relationship Inferring Network for Referring Image Segmentation

Zhiwei Hu[1†], Guang Feng[1†], Jiayu Sun[1], Lihe Zhang[1‡], Huchuan Lu[1,2]
[1]Dalian University of Technology, China [2]Peng Cheng Laboratory
hzw950822@mail.dlut.edu.cn, {fengguang.gg, jiayusun666}@gmail.com
{zhanglihe, lhchuan}@dlut.edu.cn

Abstract

Most existing methods do not explicitly formulate the mutual guidance between vision and language. In this work, we propose a bi-directional relationship inferring network (BRINet) to model the dependencies of cross-modal information. In detail, the vision-guided linguistic attention is used to learn the adaptive linguistic context corresponding to each visual region. Combining with the language-guided visual attention, a bi-directional cross-modal attention module (BCAM) is built to learn the relationship between multi-modal features. Thus, the ultimate semantic context of the target object and referring expression can be represented accurately and consistently. Moreover, a gated bi-directional fusion module (GBFM) is designed to integrate the multi-level features where a gate function is used to guide the bi-directional flow of multi-level information. Extensive experiments on four benchmark datasets demonstrate that the proposed method outperforms other state-of-the-art methods under different evaluation metrics.

1. Introduction

Referring image segmentation is a challenging task that has emerged in recent years. It helps to understand the relationship between language and vision. Unlike the simplicity of traditional semantic segmentation whose each pixel needs to be assigned a specific semantic category label, referring image segmentation requires a deeper understanding of image. In order to segment the region that best matches the referring expression, referring image segmentation needs to take into account appearance attributes, actions, spatial relationships, as well as some other cues contained in the expression. For example, if the expression is 'A man sitting on the right is wearing a black suit', we need an algorithm that not only distinguishes all the instances in the image, but also locates the most suitable one, according to the semantic meaning of the sentence.

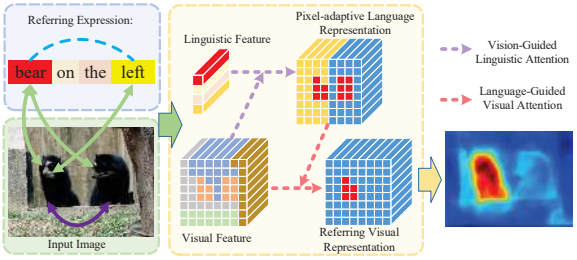

Figure 1: Bi-directional relationship inferring network. The blue dotted line represents the relationship modeling between each word. The green solid line represents the bi-directional interaction between language and vision, and the purple solid line represents the internal relationship of visual features. Given a referring expression and a query image, the bi-directional cross-modal attention module constructs the cross-modal relationship between linguistic and visual information, which makes the network pay more attention to the target object pointed by the referring expression.

Recently, the rapid development of convolutional neural network (CNN) and recurrent neural network (RNN) has greatly promoted the research progress of both computer vision and natural language processing. The most pervasive solution of referring image segmentation is to extract the visual and linguistic features by the CNN and the RNN, respectively. These features are then fused to generate the final pixel-wise segmentation mask. Some methods [14, 23, 20, 28] directly concatenate the two kinds of features and then infer the target object depending on the network itself. Due to the powerful learning ability of neural network, these methods indeed achieve some reasonable performance. However, they implicitly assume that each word contributes equally to each visual region without considering the interaction between linguistic and visual features, which often results in some inaccurate target localization. Later, some works either brutally model the relationships between the vision-word mixed features in a fully-connected manner [38] or only unidirectionally utilize the linguistic attention to formulate the cross-modal rela-

[†]Equal Contribution
[‡]Corresponding Author

tionship [31, 3]. All of them do not explicitly characterize the mutual guidance between the visual and linguistic features yet, thereby weakening the contextual consistency of linguistic and visual region in the feature space.

To this end, we propose a bi-directional relationship inferring network (BRINet) to effectively capture the dependencies of multi-modal features under the guidance of both language and vision. First, we construct a vision-guided linguistic attention module (VLAM) to learn the adaptive linguistic context for each visual region. Second, a language-guided visual attention module (LVAM) utilizes the learned linguistic context to guide the learning of spatial dependencies between any two positions of the visual features. As shown in Fig. 1, by the mutual learning between different modalities, the proposed model enriches the contextual representation of the target region. Therefore, the target region can be highlighted more consistently with the help of referring expression. This apparently allows us to consider more complex and non-sequential dependencies between visual regions and words. Finally, we design a gated bi-directional fusion module (GBFM) to guide the network in carrying out the top-down and bottom-up multi-level information aggregation selectively.

Our main contributions are listed as follows:

- We propose a novel bi-directional cross-modal attention module (BCAM) that uses both visual and linguistic guidances to capture the dependencies between multi-modal features. As a result, it can better realize the compatibility between language and visual region.
- We introduce a gated bi-directional fusion module (GBFM) as an assistant to flexibly incorporate the multi-level cross-modal features, which helps the network to further refine the segmentation result.
- BCAM and GBFM are integrated into the BRINet. Extensive experiments on four large-scale datasets show that the proposed method outperforms other state-of-the-art approaches across different metrics.

2. Related Work

2.1. Semantic Segmentation

Semantic segmentation has achieved remarkable success in recent years. Most state-of-the-art methods use the structures of fully convolutional network (FCN) [24] to generate the pixel-wise prediction in an end-to-end manner. Subsequently, many FCN-based works are proposed to alleviate the loss of details caused by the continuous downsampling and enhance the multi-scale context aggregation. PSPNet [42] utilizes a pyramid pooling module to gather some different region-based multi-scale contexts. Deeplabv2 [4] and Deeplabv3 [5] employ the atrous spatial pyramid pooling to enlarge the receptive field and embed the multi-scale contextual information. Some works [29, 1, 8] investigate the encoder-decoder structures and utilize the low-level features to complement the detailed information for more accurate prediction. Ding et al. [8] aggregate the context contrasted features to focus on the local information. Li et al. [21] use an interconnected LSTM chains to combine the multi-scale feature maps and contexts in a bi-directional and recurrent manner. DANet [11] adopts the position and channel attention to learn the spatial and channel interdependencies, respectively. Our method also considers the context combination of the different types and multiple scales of features for completely segmenting the target region.

2.2. Referring Localization and Segmentation

The goal of referring image localization is to localize an object based on a natural language expression. In [27, 15, 26], they propose a model to maximize the matching score between the target object and the given expression. Some recent methods [39, 33, 37] attempt to decompose the expression into different components and use these components to model the relationship between objects. Referring image segmentation aims to generate a precise segmentation mask instead of a bounding box for the image region by the describe of language expression. This task is first proposed in [14], which directly concatenates both visual and linguistic features to generate the final mask. In [23], a two-layered LSTM network is utilized to separately infer the tiled multi-modal feature of each visual region in a sequential manner. RRNet [20] adapts convolutional LSTM [36] to gradually fuse the pyramid features. DMNet [28] concatenates the word-specific multi-modal features in a recurrent manner. All the referring segmentation methods mentioned above follow a 'concatenation-convolution' procedure to characterize the cross-modal knowledge. However, the relationship between linguistic and visual information is not explicitly modeled. Later, KWANet [31] extracts key words to suppress the noise in the referring expression and highlight the target object. CSANet [38] designs a self-attention mechanism to model the visual attention of each word. While STEP [3] considers an image-to-word attention to compute the relevance between each word and each visual region, and also employs the resulted heatmap to recurrently guide the target segmentation. Nevertheless, these methods only realize unidirectional relationship modeling among different modalities. In this work, we have built a bi-directional guidance mechanism between linguistic and visual features so that they can better adapt to each other.

2.3. Attention Mechanism

Attention mechanism is widely applied in many tasks [32, 35, 25, 33, 43, 2, 30, 34, 41]. Deng et al. [7] introduce a co-attention mechanism to learn the adaptive

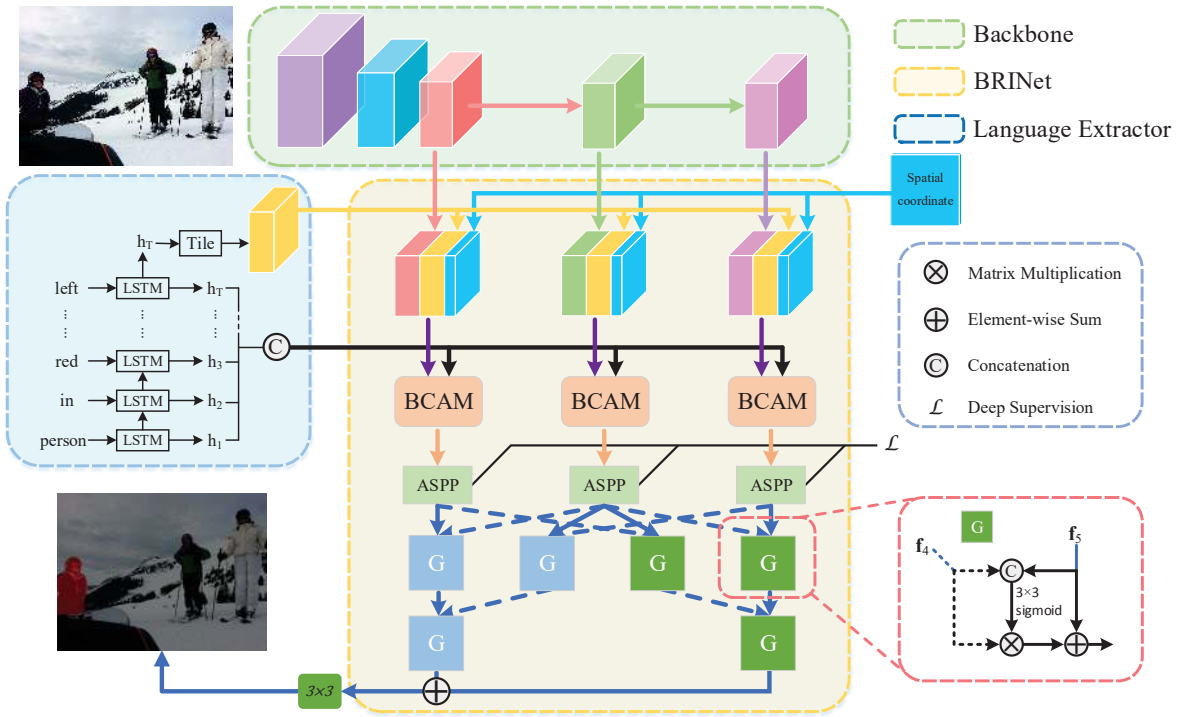

Figure 2: The overall framework of our method, where Resnet-101, shown on the top middle of the figure, encodes the feature of input image. LSTM, shown on the middle left, encodes each word in the referring expression. The bi-directional cross-modal attention module, denoted as BCAM, is used to modeling the relationship of multi-modal features. The gated bi-directional fusion module, denoted as GBFM, receives the ultimate multi-modal features from different levels to produce the final segmentation mask.

relationship between linguistic and regional features. Shi et al. [31] design a word attention to re-weight the importance between each word and image region. While Wang et al. [33] propose a graph attention to represent inter-object relationships. Yang et al. [37] construct a language-guided high-order visual relation graph among all the objects. Li et al. [19] employ a question-adaptive attention to model the multi-type relations of objects, which can learn an adaptive region representation. Different from the previous works, we extend the cross-modal attention mechanism to the task of referring image segmentation, and design a bi-directional attention to enhance the semantic consistency of feature representation.

3. The Proposed Method

The overall architecture of the proposed method is illustrated in Fig. 2. Given an image and its referring expression, we first use DeepLab ResNet-101v2 [4] and LSTM [13] to extract visual and linguistic features, respectively. Then the concatenated visual, linguistic and spatial features are fed into the bi-directional cross-modal attention module (BCAM) to model the relationships between multi-modal features. These relationships are used to update the contextual representation of the target object. Next, we use ASPP [4] to learn the multi-scale information of the updated features. Finally, the features of three levels are adaptively aggregated by the gated bi-directional fusion module (GBFM) to produce the final prediction mask.

3.1. Vision-Guided Linguistic Attention

For a given expression $L = \{l_t\}_{t=1}^{T}$, we use an LSTM [13] to represent the context of each word. The context of word l_t is denoted as $h_t \in R^{1000}$, where h_t is the hidden-state vector after running the LSTM through the first t word of L. There is a fact that the importance of each word in the sentence to the i-th feature region v_i is different. If we treat these language features equally and use them to guide image segmentation directly, some noise may be introduced to make the network produce an erroneous prediction. Thus we introduce a vision-guided linguistic attention module (VLAM) to adaptively establish the relationship between the linguistic context and each visual region. The relationship between the i-th feature region and the t-th word is defined as follows:

$$\begin{aligned} v_i^1 &= W_{v^1} v_i \\ \alpha_{i,t} &= {v_i^1}^{\mathbf{T}} h_t \\ \widetilde{\alpha}_{i,t} &= \frac{\exp(\alpha_{i,t})}{\sum_{t=1}^{T} \exp(\alpha_{i,t})}, \end{aligned} \tag{1}$$

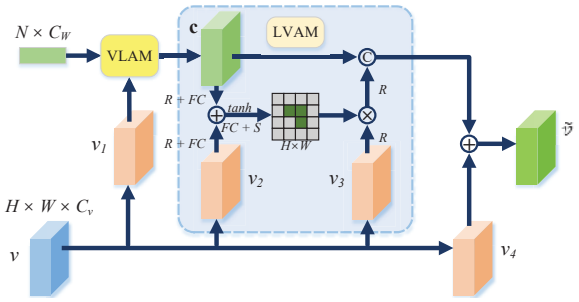

Figure 3: Bi-directional cross-modal attention module. R: Reshape; FC: Fully connected layer; S: Softmax; VLAM: Vision-guided linguistic attention module; VLAM: Language-guided visual attention module. Adaptive linguistic context **c** is used as a guide to model the relationship between input features v. With the v_i as the center, the softmax function is used to normalize the relationship weight between v_i and all feature regions.

where v_i is the concatenation of the visual feature $I_i \in R^{C_I}$, the final hidden state $h_T \in R^{1000}$ and the spatial feature $s_i \in R^8$, *i.e.*, $v_i = [I_i, h_T, s_i]$, where $[\cdot,\cdot]$ represents the concatenation operation. C_I represents the number of channels of the visual feature map. The spatial feature s_i follows the design of [14]. $W_{v^1} \in R^{1000\times(C_I+1000+8)}$ are the learnable parameters, which aim to map v_i into the same dimension of h_t. $\widetilde{\alpha}_{i,t}$ is the normalized attention score, which represents the importance of the t-th word to the i-th feature region. Thus, the new linguistic context $\mathbf{c}_i$ for the i-th feature region can be calculated as follows:

$$\mathbf{c}_i = \sum_{t=1}^{T} \widetilde{\alpha}_{i,t} h_t. \tag{2}$$

3.2. Language-Guided Visual Attention

Contextual information is essential for referring image segmentation, which helps the network locate and segment the object region accurately. To model contextual relationship across different regions, we design a language-guided visual attention module (LVAM) that leverages the region-adaptive linguistic features to compute their affinities.

For feature vector v_i, the normalized relationship weighted between itself and j-th region v_j is defined as follows:

$$\begin{aligned} v_j^2 &= W_{v^2} v_j \\ \lambda_{i,j} &= W_\lambda[\tanh(W_c \mathbf{c}_i + W_{\widetilde{v}^2} v_j^2)] \\ \widetilde{\lambda}_{i,j} &= \frac{\exp(\lambda_{i,j})}{\sum_{j=1}^{N}\exp(\lambda_{i,j})}, \end{aligned} \tag{3}$$

where $W_{v^2} \in R^{1000\times(C_I+1000+8)}$, $W_c \in R^{500\times1000}$, $W_{\widetilde{v}^2} \in R^{500\times1000}$ and $W_\lambda \in R^{1000}$ are the learnable parameters. $\mathbf{c}_i$ defined in Eq. 2,N is the number of pixels. $\widetilde{\lambda}_{i,j}$ is the importance of the j-th feature towards the i-th feature region. Based on the this process, we establish the dependencies of all regions in the image. Therefore, on the next stage, we use these relationships to update the visual feature representation,

$$\begin{aligned} v_j^3 &= W_{v^3} v_j, \quad v_i^4 = W_{v^4} v_i \\ \widetilde{v}_i &= W_{\widetilde{v}}[\sum_{j=1}^{N}(\widetilde{\lambda}_{i,j} v_j^3),\ \mathbf{c}_i] + v_i^4, \end{aligned} \tag{4}$$

where W_{v^3}, $W_{v^4} \in R^{1000\times(C_I+1000+8)}$ and $W_{\widetilde{v}} \in R^{1000\times2000}$ are the learnable parameters. Fig. 3 shows the detailed structures of VLAM and LVAM, which constitute the cross-modal attention module.

3.3. Gated Bi-directional Fusion

Previous works [29, 1, 8] on semantic segmentation demonstrate that the encoder-decoder structure can integrate the multi-level features to further refine the segmentation mask. Inspired by them, we introduce a gated bi-directional fusion module (GBFM), which detailed architecture is shown in Fig. 2. We define the output of ASPP as $\mathbf{F} = \{\mathbf{f}_i\}_{i=3}^{5}$, which corresponding to $Res3$, $Res4$ and $Res5$, respectively. $\{\mathbf{f}_i\}_{i=3}^{5}$ have the same channel number and resolution. We use both bottom-up and top-down manners to guide the multi-level feature fusion gradually.

In the bottom-up pathway, we expect that the higher-level features provide the global and semantic guidance to the lower-level ones. The process is shown as follows:

$$\begin{aligned} \mathbf{f}_{34}^U = G_{3,4}^U \otimes \mathbf{f}_3 + \mathbf{f}_4, \quad \mathbf{f}_{45}^U = G_{4,5}^U \otimes \mathbf{f}_4 + \mathbf{f}_5 \\ \mathbf{f}_{\text{fuse}}^U = G_{34,45}^U \otimes \mathbf{f}_{34}^U + \mathbf{f}_{45}^U, \end{aligned} \tag{5}$$

where $\otimes$ is the element-wise product. G^U is the gate function, which is used to control the information flow. The gate function can be calculated as follows:

$$G_{i,j}^U = Sig(Conv(Cat(\mathbf{f}_i, \mathbf{f}_j))), \tag{6}$$

where $Cat(\cdot,\cdot)$ represents the concatenation operation along the channel axis. $Conv$ denotes a 3×3 convolutional layer. Sig denotes the element-wise sigmoid function.

In the top-down pathway, we hope that the lower-level features provide the local and fine guidance to the higher-level ones. The process is shown as follows:

$$\mathbf{f}_{\text{fuse}}^D = (\mathbf{f}_3 + G_{3,4}^D \otimes \mathbf{f}_4) + G_{34,45}^D \otimes (\mathbf{f}_4 + G_{4,5}^D \otimes \mathbf{f}_5), \tag{7}$$

where $\otimes$ and G^D have the same meaning as the symbols in Eq. 5. Similarly, the gate function can be calculated using

*	ReferIt	UNC			UNC+			G-Ref
	test	val	testA	testB	val	testA	testB	val
LSTM-CNN [14]	48.03	-	-	-	-	-	-	28.14
RMI+DCRF [23]	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
DMN [28]	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
KWA [31]	59.19	-	-	-	-	-	-	36.92
RRN+DCRF [20]	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
MAttNet [39]	-	56.51	62.37	51.70	46.67	52.39	40.08	-
lang2seg [6]	-	58.90	61.77	53.81	-	-	-	-
CMSA+DCRF [38]	63.80	58.32	60.61	55.09	43.76	47.60	37.89	39.98
STEP [3]	64.13	60.04	63.46	57.97	48.19	52.33	40.41	46.40
Ours	63.11	60.98	62.99	59.21	48.17	52.32	42.11	47.57
Ours+DCRF	63.46	61.35	63.37	59.57	48.57	52.87	42.13	48.04

Table 1: Quantitative results of overall IoU on four datasets.'-' denotes no available results. DCRF means DenseCRF [18] post-processing.

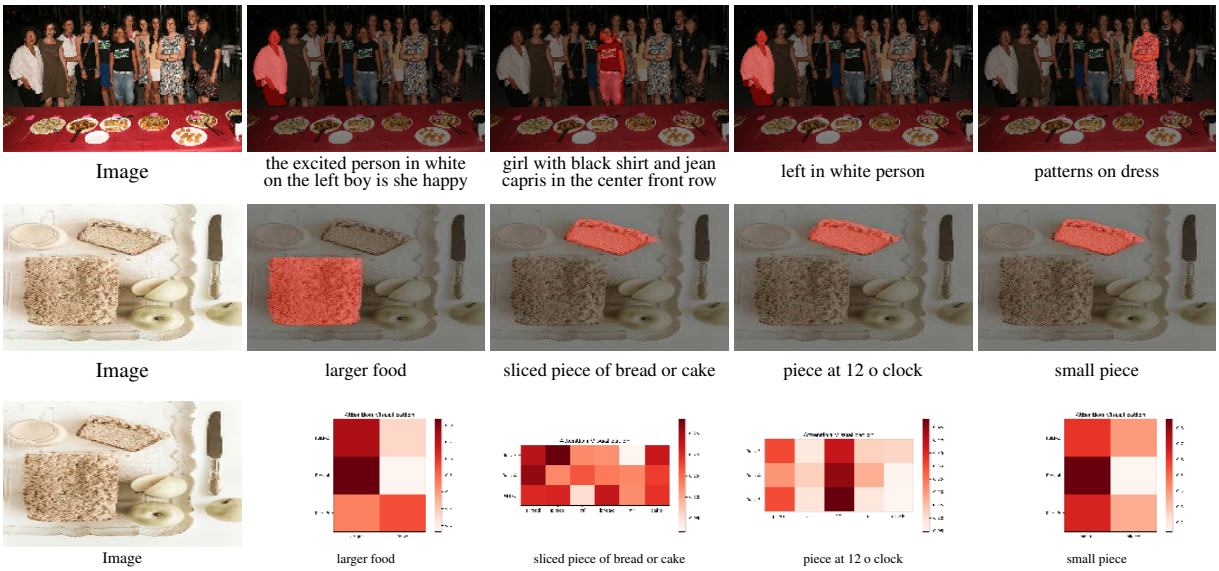

Figure 4: Visual examples of referring image segmentation by the BRINet.

$G^D_{i,j} = Sig(Conv(Cat(\mathbf{f}_i, \mathbf{f}_j)))$. Finally, the fusion feature can be obtained by

$$\mathbf{f}_{\text{final}} = Conv(\mathbf{f}^U_{\text{fuse}} + \mathbf{f}^D_{\text{fuse}}), \tag{8}$$

which is used to calculated the final prediction.

4. Experiments

4.1. Datasets

To evaluate the performance of our model, we use four referring image segmentation datasets: UNC [40], UNC+ [40], Google-Ref [27] and ReferIt [16].

UNC: The UNC dataset is based on MS COCO [22] dataset using a two-player game [16]. It contains 19,994 images with 142,209 expressions referring for 50,000 segmented image regions. More than one object of the same category appears in each image.

UNC+: The UNC+ dataset consists of 141,564 language expressions for 49,856 objects in 19,992 images. Similar to the UNC dataset, its images and referring expressions are also selected from MS COCO [22]. However, there is a restriction of this dataset that no words in the referring expressions indicate location. Namely, the expression of the objects only describe the appearance information.

Google-Ref: Google-Ref is built on top of MS COCO [22] dataset. There are 104,560 expressions referring to 54,822 objects in 26,711 images. All annotations of this dataset are collected on Mechanical Turk instead of using a two-player game. Each image contains 2 to 4 objects of the same category, and the average length of referring expressions is 8.43 words. Thus, the referring expressions are longer and the descriptions are richer.

ReferIt: The ReferIt dataset is composed of 130,525 expressions referring to 96,654 object regions in 19,894 im-

	Method	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	overall IoU
val	Baseline	56.50	48.31	37.55	23.21	4.95	52.26
	BCAM w/o language	59.23	51.36	40.95	26.35	6.15	53.91
	BCAM w/o VLAM	62.47	54.41	42.96	27.00	6.03	55.14
	BCAM	65.53	57.46	46.85	30.42	7.28	56.76
	BRINet w/o Gate	69.54	62.51	52.57	35.35	9.91	59.59
	BRINet w/ left	70.89	64.27	54.62	37.62	10.51	60.32
	BRINet w/ right	71.20	64.36	55.22	38.16	10.65	60.55
	BRINet	71.83	65.05	55.64	39.36	11.21	61.35
	$\mathbf{f}_4 + \mathbf{f}_5$	70.69	63.83	53.61	36.81	9.78	60.09
	ConvLSTM	69.61	62.67	52.82	36.26	9.34	58.78
	BRINet w/o SP	67.81	60.84	51.08	65.48	9.78	58.91
testA	Baseline	59.61	51.00	40.39	25.76	5.69	54.24
	BCAM w/o language	62.52	55.10	44.79	29.34	6.77	55.52
	BCAM w/o VLAM	65.12	55.65	45.27	29.40	6.20	56.40
	BCAM	68.84	61.06	49.96	32.01	7.55	59.04
	BRINet w/o Gate	72.54	65.58	54.75	36.43	9.42	61.50
	BRINet w/ left	74.70	68.08	57.52	39.19	9.97	62.80
	BRINet w/ right	75.08	68.59	58.55	40.09	10.29	62.99
	BRINet	75.09	68.29	58.37	41.01	10.96	63.37
	$\mathbf{f}_4 + \mathbf{f}_5$	74.19	67.46	57.38	38.22	9.83	62.25
	ConvLSTM	72.67	65.83	55.54	37.97	9.60	60.75
	BRINet w/o SP	71.95	65.15	54.96	37.94	9.65	61.95
testB	Baseline	54.09	45.69	35.84	23.00	6.73	50.84
	BCAM w/o language	56.29	48.24	38.21	25.20	7.73	52.28
	BCAM w/o VLAM	57.64	49.50	39.57	25.91	7.99	52.17
	BCAM	61.00	52.38	42.43	28.64	9.36	54.12
	BRINet w/o Gate	66.24	58.63	49.64	35.64	12.60	57.81
	BRINet w/ left	67.05	59.84	51.13	37.33	13.17	58.13
	BRINet w/ right	66.95	59.71	50.79	37.31	13.76	58.50
	BRINet	68.38	61.77	52.76	38.14	14.33	59.57
	$\mathbf{f}_4 + \mathbf{f}_5$	66.99	60.08	50.64	35.90	13.19	57.91
	ConvLSTM	65.77	58.33	49.74	35.56	12.99	56.56
	BRINet w/o SP	63.32	55.76	46.61	33.46	11.72	56.18

Table 2: Ablation study on the UNC val, testA and testB datasets.

ages. ReferIt is collected from the IAPR TC-12 [9] dataset. The foreground regions consist of objects and stuff (e.g., ground, mountain and sky), and the expressions are usually shorter and more succinct than the other datasets.

4.2. Implementation Details

Given an input image, we resize and zero-pad it to 320×320. The DeepLab ResNet-101v2 [4] is used as visual feature extractor in this work. Similar to all previous methods, this network is pretrained on the Pascal VOC dataset [10]. This is because that all previous methods pretrain their models on the Pascal VOC. We use the outputs of the DeepLab blocks $Res3$, $Res4$ and $Res5$ as the inputs for our module and use the notations v_i ($i \in 3, 4, 5$) to denote the corresponding features. The resolution of each feature map v_i is 40×40.

Following [23, 20], the size of each LSTM cell is set to 1000. The maximum length of language expression is 20. In other words, we keep only the first 20 words of each expression. This is because most of the language expressions on the benchmark datasets are shorter than the predefined maximum length, which ensures the integrity of the input sentence in most cases. Similar to the implementation of [14], we concatenate an 8-D spatial coordinate feature to further enhance the spatial information of v_i.

Our network is trained by an end-to-end strategy and chooses Adam [17] optimizer with an initial learning rate of 0.00025. The weight decay and batch size are 0.0005 and 1, respectively. The initial learning rate gradually decreases by a polynomial decay with power of 0.9. For a fair comparison, all the final predicted segmentation masks are refined by DenseCRF [18].

Metrics: Following the previous work [20, 23, 31, 38], we use two typical metrics to evaluate the segmentation accuracy: Overall Intersection-over-Union (Overall IoU) and Precision@X. The Overall IoU metric calculates the ratio

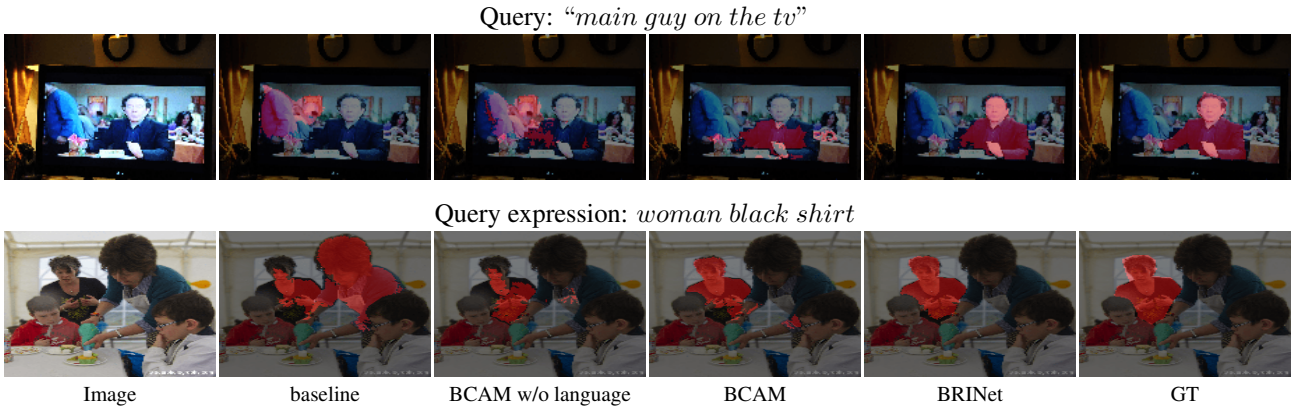

Figure 5: Visual examples of the proposed modules.

	Length	1-5	6-7	8-10	11-20
G-Ref	R+LSTM [23]	32.29	28.27	27.33	26.61
	R+RMI [23]	35.34	31.76	30.66	30.56
	Ours	51.93	47.55	46.33	46.49
	Length	1-2	3	4-5	6-20
UNC	R+LSTM [23]	43.66	40.60	33.98	24.91
	R+RMI [23]	44.51	41.86	35.05	25.95
	Ours	65.99	64.83	56.97	45.65
	Length	1-2	3	4-5	6-20
UNC+	R+LSTM [23]	34.40	24.04	19.31	12.30
	R+RMI [23]	35.72	25.41	21.73	14.37
	Ours	59.12	46.89	40.57	31.32
	Length	1	2	3-4	5-20
ReferIt	R+LSTM [23]	67.64	52.26	44.87	33.81
	R+RMI [23]	68.11	52.73	45.69	34.53
	Ours	75.28	62.62	56.14	44.40

Table 3: IoU for different length referring expressions on Google-Ref, UNC, UNC+ and ReferItGame.

of the total intersection regions and the total union regions between the predicted mask and the ground truth. The second metric calculates the percentage of the images with IoU higher than the threshold X during the testing process, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.3. Performance Comparison

We compare the proposed method with some state-of-the-art methods, including LSTM-CNN [14], RMI [23], DMN [28], KWA [31], RRN [20], MAttNet [39], lang2seg [6] CMSA [38], and STEP [3].

Quantitative Evaluation: The segmentation performance (IoU) of these methods on all datasets is summarized in Tab. 1. We observe that the proposed method outperforms the other approaches across different datasets, especially on the G-Ref. G-Ref is more complicated than the other datasets, because of its longer expressions for object referral. Our method outperforms the second best by 20.16%. It is worth noting that the two methods, MAttNet and lang2seg, use Mask R-CNN [12] to pre-process and post-process images when segmenting images. Mask R-CNN itself can better locate and segment all targets in the image, which obviously contributes much to the performance improvement. This actually indicates that the designed end-to-end method has significant performance advantage over these methods. In addition, the referring expression in the UNC+ dataset does not include the words that indicate spatial or location information, which puts forward higher demand upon the comprehension ability of the appearance of objects. The significant improvement on the UNC+ dataset shows that the BRINet can more comprehensively understand the semantics of objects.

We further study the relationship between the segmentation performance and the referring expressions length. We divide the expression into four groups according to [23], and the segmentation results of each group are shown in Tab. 3. The BRINet outperforms the other methods on all groups.

Qualitative Evaluation: Fig. 4 shows some representative results and visualization examples of linguistic attention to exhibit the superiority of the proposed method. It can be seen that our method can accurately segment the target object region, even when the lengths of referring expression are various and the scenes are complex. Besides, Our method is also robust to the referring image segmentation without location or spatial information in referring expression (row 2).

4.4. Ablation Study

The proposed framework is mainly composed of two modules, including BCAM and GBFM. To further investigate the relative contribution of each components in BRINet, we conduct a series of experiments on the UNC

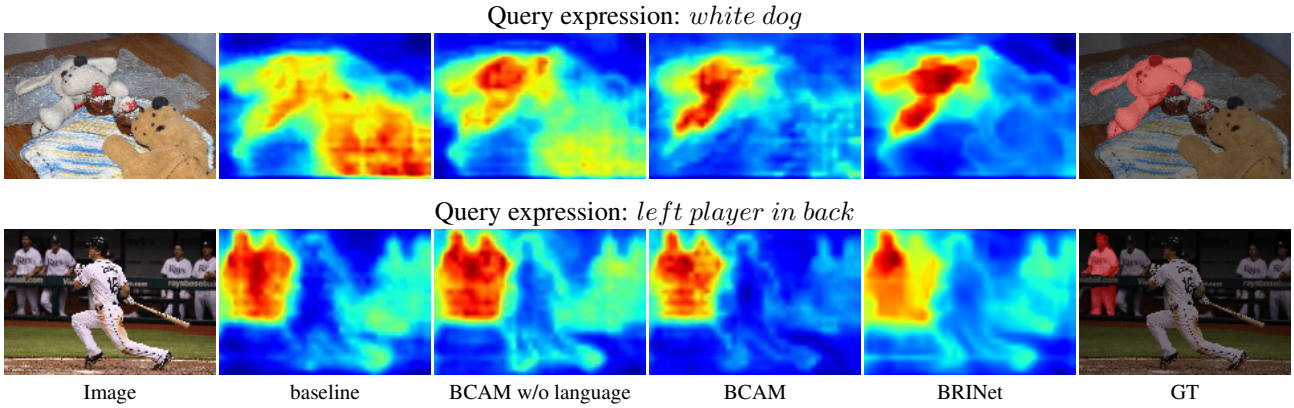

Figure 6: Segmentation heatmaps of the proposed modules.

dataset. We also verify the impact of 8-D spatial coordinates on performance (BRINet w/o SP). The detailed experimental results are shown in Tab. 2.

Effectiveness of BCAM: We remove BCAM and G-BFM from the BRINet shown in Fig. 2 to build the baseline network. In Tab. 2, we gradually analyze BCAM without language guidance (BCAM w/o language), BCAM without VLAM (BCAM w/o VLAM) and the complete B-CAM schemes, whose mechanisms realise the visual self-attention, the plain language-guided relationship inference and the adaptive language-guided relationship inference respectively. BCAM w/o language scheme indicates that the relationship modeling between features is beneficial to enhance the segmentation results. Experimental comparisons between BCAM w/o VLAM and the complete BCAM further verify that the adaptive linguistic features are beneficial to learn the relationship between visual features.

Effectiveness of GBFM: The gated bi-directional fusion module (GBFM) consists of two components: the bottom-up and top-down information fusion modules. Here, we compare the results of the direct summation of multi-level features (BRINet w/o Gate), the top-down fusion (BRINet w/ left), the bottom-up fusion (BRINet w/ right) and the full BRINet on UNC dataset. As shown in Tab. 2, we can find that both top-down and bottom-up message passing models controlled by the gate function can effectively improve performance and the full BRINet achieves the best results. In addition, we analyze the influence of the number of scales, where 'BRINet w/o Gate' means that only feature $\mathbf{f}_5$ is used, while 'BRINet' adopts three scales $\mathbf{f}_3 + \mathbf{f}_4 + \mathbf{f}_5$. The results of $\mathbf{f}_4 + \mathbf{f}_5$ are also given.

We compare GBFM with ConvLSTM in Tab. 2. The latter formulates the input-to-state and state-to-state transitions. It mainly models the long-range dependancies via the cascaded structure. While GBFM aims to model multi-level feature fusion, which utilizes two pyramidally stacked structures to achieve the vertically skipping layer fusion and the horizontally bi-directional information fusion. Moreover, the latter requires much more complicated computation and parameters (4 times) than the former.

Some visual results in Fig. 5 and Fig. 6 demonstrate the benefits of each module. Among them, the visualization heatmaps are generated by the same technique as in [20, 38], which normalizes the strongest activated channel of the last feature map and up-samples it back to the same size of the input image. These figures show that the guidance mechanism can help achieve high-level contextual consistency between the referring expression and the target region. By the mutual guidance of BCAM and the feature refinement of GBFM, the network eliminates the influence of ambiguous objects and produces a good result.

5. Conclusion

In this paper, we propose a novel bi-directional relationship inferring network (BRINet) for referring image segmentation. It consists of the bi-directional cross-modal attention module (BCAM) and the gated bi-directional fusion module (GBFM). BCAM realizes the mutual guidance between linguistic and visual features, which encourages the accurate and consistent semantic representations between the referring expression and the target object. GBFM is used to adaptively filter information between features of different levels. The gate can control the information flow to better integrate multi-level cues. The experimental results on four datasets demonstrate that the proposed method achieves state-of-the-art performance.

Acknowledgements

This work was supported in part by the National Key R&D Program of China #2018AAA0102003, National Natural Science Foundation of China #61876202, #61725202, #61751212 and #61829102, and the Dalian Science and Technology Innovation Foundation #2019J12GX039.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017.

[2] Boyu Chen, Peixia Li, Chong Sun, Dong Wang, Gang Yang, and Huchuan Lu. Multi attention module for visual tracking. *Pattern Recognition*, 87:80–93, 2019.

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

[4] 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 TPAMI*, 40(4):834–848, 2017.

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[6] Yi-Wen Chen, Yi-Hsuan Tsai, Tiantian Wang, Yen-Yu Lin, and Ming-Hsuan Yang. Referring expression object segmentation with caption-aware consistency. *BMVC*, 2019.

[7] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *CVPR*, pages 7746–7755, 2018.

[8] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, pages 2393–2402, 2018.

[9] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. The segmented and annotated iapr tc-12 benchmark. *Computer vision and image understanding*, 114(4):419–428, 2010.

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

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

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124. Springer, 2016.

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

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

[18] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011.

[19] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, October 2019.

[20] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, pages 5745–5753, 2018.

[21] Di Lin, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Multi-scale context intertwining for semantic segmentation. In *ECCV*, pages 603–619, 2018.

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

[23] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, pages 1271–1280, 2017.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[25] Yifan Lu, Jiaming Lu, Songhai Zhang, and Peter Hall. Traffic signal detection and classification in street views using an attention model. *Computational Visual Media*, 4(3):253–266, 2018.

[26] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, pages 7102–7111, 2017.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[28] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, pages 630–645, 2018.

[29] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015.

[30] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *AAAI*, 2018.

[31] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, pages 38–54, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[33] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Refer-

ring expression comprehension via language-guided graph attention networks. In *CVPR*, pages 1960–1968, 2019.

[34] Tiantian Wang, Yongri Piao, Xiao Li, Lihe Zhang, and Huchuan Lu. Deep learning for light field saliency detection. In *ICCV*, pages 8838–8848, 2019.

[35] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018.

[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 *NeurIPS*, pages 802–810, 2015.

[37] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *CVPR*, pages 4145–4154, 2019.

[38] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019.

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

[40] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016.

[41] Lihe Zhang, Jie Wu, Tiantian Wang, Ali Borji, Guohua Wei, and Huchuan Lu. A multistage refinement network for salient object detection. *IEEE TIP*, 29:3534–3545, 2020.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017.

[43] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, pages 267–283, 2018.