Linguistic Structure Guided Context Modeling for Referring Image Segmentation

Tianrui Hui[1,2], Si Liu[3](✉), Shaofei Huang[1,2], Guanbin Li[4], Sansi Yu[5], Faxi Zhang[5], and Jizhong Han[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
{huitianrui,huangshaofei,hanjizhong}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Institute of Artificial Intelligence, Beihang University, Beijing, China
liusi@buaa.edu.cn
[4] Sun Yat-sen University, Guangzhou, China
liguanbin@mail.sysu.edu.cn
[5] Tencent Marketing Solution, Shenzhen, China
{mionyu,micahzhang}@tencent.com

Abstract. Referring image segmentation aims to predict the foreground mask of the object referred by a natural language sentence. Multimodal context of the sentence is crucial to distinguish the referent from the background. Existing methods either insufficiently or redundantly model the multimodal context. To tackle this problem, we propose a "gather-propagate-distribute" scheme to model multimodal context by cross-modal interaction and implement this scheme as a novel Linguistic Structure guided Context Modeling (LSCM) module. Our LSCM module builds a Dependency Parsing Tree suppressed Word Graph (DPT-WG) which guides all the words to include valid multimodal context of the sentence while excluding disturbing ones through three steps over the multimodal feature, i.e., gathering, constrained propagation and distributing. Extensive experiments on four benchmarks demonstrate that our method outperforms all the previous state-of-the-arts.

Keywords: Referring segmentation · Multimodal context · Linguistic structure · Graph propagation · Dependency Parsing Tree

1 Introduction

Referring image segmentation aims at predicting the foreground mask of the object which is matched with the description of a natural language expression. It

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58607-2_4) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12355, pp. 59–75, 2020.
https://doi.org/10.1007/978-3-030-58607-2_4

enjoys a wide range of applications, e.g., human-computer interaction and interactive image editing. Since natural language expressions may contain diverse linguistic concepts, such as entities (e.g. “car”, “man”), attributes (e.g. “red”, “small”) and relationships (e.g. “front”, “left”), this task is faced with a broader set of categories compared with a predefined one in traditional semantic segmentation. It requires the algorithm to handle the alignment of different semantic concepts between language and vision.

A general solution to this task is first extracting visual and linguistic features respectively, and then conducting segmentation based on the multimodal features generated from the two types of features. The entity referred by a sentence is defined as the *referent*. Multimodal features of the referent is hard to be distinguished from features of the background due to the existence of abundant noises. To solve this problem, valid multimodal context relevant to the sentence can be exploited to highlight features of the referent and suppress those of the background for accurate segmentation. Some works tackle this problem by straightforward concatenation [16,32] or recurrent refinement [4,21,25] of visual and linguistic features but lack the explicit modeling of multimodal context. Other works introduce dynamic filters [29] or cross-modal self-attention [39] to model multimodal context. However, these multimodal contexts are either insufficient or redundant since the number of dynamic filters [29] is limited and weights for aggregating multimodal context in self-attention [39] may be redundant due to dense computation operations.

To obtain valid multimodal context, a feasible solution is to exploit linguistic structure as guidance to selectively model valid multimodal context which is relevant to the sentence. As illustrated in Fig. 1, each word can gather multimodal context related to itself by cross-modal attention. For example, the word “dog” corresponds to the red masks of two dogs in the image. Multimodal context of each word is a partial and isolated comprehension result of the whole sentence. Therefore, constrained communication among words is required to include valid multimodal context and exclude disturbing ones. Afterwards, communicated multimodal context of each word contains appropriate information relevant to the whole sentence and can be aggregated to form valid multimodal context for highlighting features of the referent.

To realize the above solution, we propose a Linguistic Structure guided multimodal Context Modeling (LSCM) module in this paper. Concretely, features of the input sentence and image are first fused to form the multimodal features. Then, as illustrated in Fig. 1, in order to fully exploit the linguistic structure of the input sentence, we construct a Dependency Parsing Tree suppressed Word Graph (DPT-WG) where each node corresponds to a word. Based on the DPT-WG, three steps are conducted to model valid multimodal context of the sentence. (1) **Gather** relevant multimodal features (i.e., context) corresponding to a specific word through cross-modal attention as the node feature. At this step, each word node contains only multimodal context related to itself. Take Fig. 1 as an example, the segments corresponding to “dog” and “table” are denoted by red and blue masks respectively. The multimodal features inside each mask are attentively gathered to form the node feature of the graph. (2) **Propagate** information

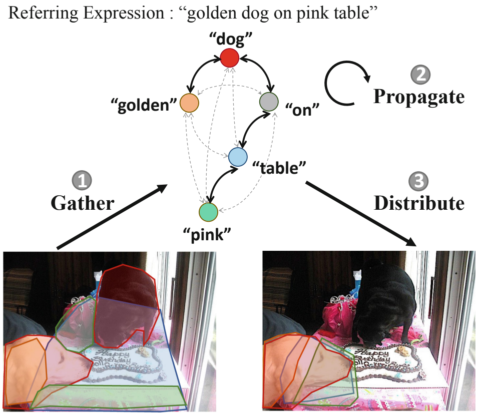

Fig. 1. Illustration of our proposed LSCM module. We construct a Dependency Parsing Tree suppressed Word Graph (DPT-WG) to model multimodal context in three steps. 1) **Gather.** Multimodal context relevant to each word are gathered as feature of each word node. Therefore, each word corresponds to some visually relevant segments in the image. For example, word “dog” corresponds to two red segments in the left image. 2) **Propagate.** DPT is exploited to further guide each word node to include valid multimodal context from others and exclude disturbing ones through suppressed graph propagation routes. Gray dotted and black solid lines denote suppressed and unsuppressed edges in DPT-WG respectively. 3) **Distribute.** Features of all word nodes are distributed back to the image. Segments corresponding to the input words are all clustered around the ground-truth segmentation region, i.e., the golden dog on pink table in the right image. (Best viewed in color).

among word nodes so that each word node can obtain multimodal context of the whole sentence. Initially, nodes in the word graph are fully-connected without any constraint on the edge weights. However, two words in the sentence may not be closely relevant to each other and unconstrained communication between them may introduce disturbing multimodal context. For example, the words “golden” and “pink” in Fig. 1 modify different entities respectively (“dog” and “table”) and have relatively weak relevance between each other. Unconstrained (i.e., extensive) information propagation between “golden” and “pink” is unnecessary and may introduce disturbing multimodal context. Therefore, we utilize Dependency Parsing Tree (DPT) [3] to describe syntactic structures among words to selectively suppress certain weights of edges in our word graph. The DPT-WG can guide each word node to include valid contexts from others and exclude disturbing ones. After propagation, updated node features acquire information of the whole sentence. As shown in Fig. 1, the five words communicate and update their features under the structural guidance of our DPT-WG. (3) **Distribute** the updated node features back to every spatial location on the multimodal feature map. As shown in Fig. 1, the segments corresponding to the input words are all clustered around the ground-truth referring segmentation. It shows the updated multimodal features contain more valid multimodal context. In addition, we also propose a Dual-Path

Multi-Level Fusion module which integrates spatial details of low-level features and semantic information of high-level features using bottom-up and top-down paths to refine segmentation results.

The main contributions of our paper are summarized as follows:

– We introduce a "gather-propagate-distribute" scheme to model compact multimodal context by interaction between visual and linguistic modalities.
– We implement the above scheme by proposing a Linguistic Structure guided Context Modeling (LSCM) module which can aggregate valid multimodal context and exclude disturbing ones under the guidance of Dependency Parsing Tree suppressed Word Graph (DPT-WG). Thus, more discriminative multimodal features of the referent are obtained.
– Extensive experiments on four benchmarks demonstrate that our method outperforms all the previous state-of-the-arts, i.e., UNC (+1.58%), UNC+ (+3.09%), G-Ref (+1.65%) and ReferIt (+2.44%).

2 Related Work

2.1 Semantic Segmentation

In recent years, semantic segmentation has made great progress with Fully Convolutional Network [27] based methods. DeepLab [5] replaces standard convolution with atrous convolution to enlarge the receptive field of filters, leading to larger feature maps with richer semantic information than original FCN. DeepLab v2 [6] and v3 [7] employ parallel atrous convolutions with different atrous rates called ASPP to aggregate multi-scale context. PSPNet [43] adopts a pyramid pooling module to capture multi-scale information. EncNet [42] encodes semantic category prior information of the scenes to provide global context. Many works [1,23] exploit low level features containing detailed information to refine local parts of segmentation results.

2.2 Referring Image Localization and Segmentation

Referring image localization aims to localize the object referred by a natural language expression with a bounding box. Some works [15,22,36] model the relationships between multimodal features to match the objects with the expression. MAttNet [40] decomposes the referring expression into subject, location and relationship to compute modular scores for localizing the referent. Comparing with referring image localization, referring image segmentation aims to obtain a more accurate result of the referred object, i.e., a semantic mask instead of a bounding box. Methods in the referring segmentation field can be divided into two types, i.e., bottom-up and top-down. **Bottom-up** methods mainly focus on multimodal feature fusion to directly predict the mask of the referent. Hu *et al.* [16] proposes a straightforward concatenation of visual and linguistic features from CNN and LSTM [13]. Multi-level feature fusion are exploited in [21]. Word attention [4,32], multimodal LSTM [25,29] and adversarial learning [31] are further incorporated to refine multimodal features. Cross-modal self-attention is exploited in [39] to capture the long-range dependencies between image regions and words, introducing much redundant context due to the dense computation of self-attention. **Top-down** methods mainly rely on pretrained pixel-level detectors, i.e., Mask R-CNN [11] to generate RoI proposals and predict the mask within the selected proposal. MAttNet [40] incorporates modular scores into Mask R-CNN framework to conduct referring segmentation task. Recent CAC [8] introduces cycle-consistency between referring expression and its reconstructed caption into Mask R-CNN to boost the segmentation performance. In this paper, we propose a bottom-up method which exploits linguistic structure as guidance to include valid multimodal context and exclude disturbing ones for accurate referring segmentation.

2.3 Structural Context Modeling

Modeling context information is vital to vision and language problems. Typical methods like self-attention [33,34] has shown great power for capturing the long range dependencies within the linguistic or visual modality. In addition, more complicated data structures are also explored to model context information. Chen *et al.* [9] proposes a latent graph with a small number of nodes to capture context from visual features for recognition and segmentation. In referring expression task, graphs [14,36–38] using region proposals as nodes and neural module tree traversal [26] are also explored to model multimodal contexts to some extent. Different from them, we propose to build a more compact graph using referring words as nodes and exploit dependency parsing tree [3] to selectively model valid multimodal context.

3 Method

The overall architecture of our model is illustrated in Fig. 2. We first extract visual and linguistic features with a CNN and an LSTM respectively and then fuse them to obtain the multimodal feature. Afterwards, the multimodal feature is fed into our proposed Linguistic Structure guided Context Modeling (LSCM) module to highlight multimodal features of the referred entity. Our LSCM module conducts context modeling over the multimodal features under the structural guidance of DPT-WG. Finally, multi-level features are fused by our proposed Dual-Path Fusion module for mask prediction.

3.1 Multimodal Feature Extraction

Our model takes an image and a referring sentence with T words as input. As shown in Fig. 2, we use a CNN backbone to extract multi-level visual features and then transform them to the same size. Multi-level visual features $\{V_2, V_3, V_4, V_5\}$ correspond to $\{Res2, Res3, Res4, Res5\}$ features of ResNet [12], where $V_i \in \mathbb{R}^{H\times W\times C_v}, i \in \{2, 3, 4, 5\}$, with H, W and C_v being the height, width and channel number of visual features respectively. Since we conduct the

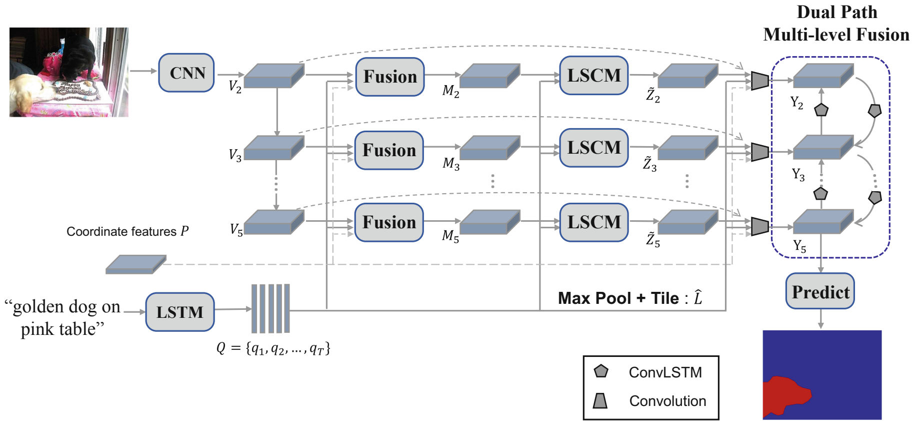

Fig. 2. Overall architecture of our model. Multi-level visual features $V_i, i \in [2, 5]$, word features Q and coordinate feature P are first fused to get multimodal features M_i. Then M_i are fed into our proposed LSCM to model valid multimodal context guided by linguistic structures. The output features $\tilde{Z}_i$ are combined with previous features and further fused through our Dual-Path Multi-Level Fusion module for mask prediction.

same operations on each level of the visual features, we use V to denote a single level of them for ease of presentation. For the input sentence of T words, we generate features of all the words $Q \in \mathbb{R}^{T \times C_l}$ with an LSTM [13]. To incorporate more spatial information, we also use an 8D spatial coordinate feature [25] denoted as $P \in \mathbb{R}^{H \times W \times 8}$. Afterwards, we fuse the features $\{V, Q, P\}$ to form the multimodal feature $M \in \mathbb{R}^{H \times W \times C_h}$, for which a simplified Mutan fusion [2] is adopted in this paper: $M = Mutan(V, Q, P)$. Details of Mutan fusion are included in the supplementary materials. Note that our method is not restricted to Mutan fusion, any other multimodal fusion approach can be used here.

3.2 Linguistic Structure Guided Context Modeling

In this module, we build a Dependency Parsing Tree suppressed Word Graph (DPT-WG) to model valid multimodal context. As illustrated in Fig. 3, we first gather feature vectors of all the spatial locations on multimodal feature M into T word nodes of WG. Then we exploit DPT [3] to softly suppress the disturbing edges in WG for selectively propagating information among word nodes, which includes valid multimodal contexts while excluding disturbing ones. Finally, we distribute features of word nodes back to each spatial location.

Gather: We get a cross-modal attention map $B \in \mathbb{R}^{T \times HW}$ with necessary reshape and transpose operations as follows:

$$B' = (QW_{q2})(MW_m)^T, \tag{1}$$

$$B = Softmax(\frac{B'}{\sqrt{C_h}}), \tag{2}$$

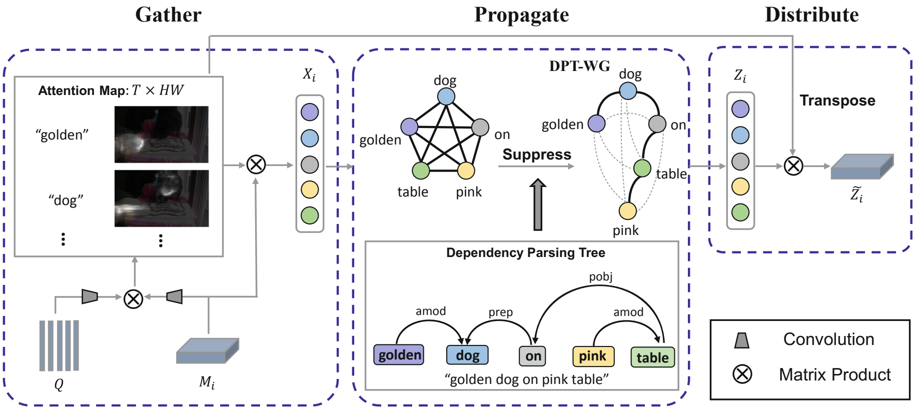

Fig. 3. Illustration of our LSCM module. We use cross-modal attention between words features Q and multimodal feature M_i to gather feature for each word node. Then we exploit DPT to softly suppress disturbing edges in the initial fully-connected WG and conduct information propagation. Finally, the updated features of word nodes are distributed back as $\tilde{Z}_i$ to incorporate valid multimodal context into original features.

where $W_{q2} \in \mathbb{R}^{C_l \times C_h}$ and $W_m \in \mathbb{R}^{C_h \times C_h}$ are learned parameters. Then we apply the normalized attention map B to M to gather the features into T word nodes:

$$X = BM, \tag{3}$$

where $X = [x_1; x_2; ...; x_T] \in \mathbb{R}^{T \times C_h}$ denotes the features of word nodes. Each $x_t, t = 1, 2, ..., T$ encodes the multimodal context related with the t-th word.

Propagate: The word graph used for context modeling is fully-connected. Thus, the adjacency matrix $A \in \mathbb{R}^{T \times T}$ is computed as follows:

$$A' = (XW_{x1})(XW_{x2})^T, \tag{4}$$

$$A = Softmax(\frac{A'}{\sqrt{C_h}}), \tag{5}$$

where $W_{x1} \in \mathbb{R}^{C_h \times C_h}$, $W_{x2} \in \mathbb{R}^{C_h \times C_h}$ are parameters for linear transformation layers. At present, the edge weights among word nodes are represented by multimodal feature similarities which are unconstrained. However, two words may not be closely related in the sentence and unconstrained information propagation between them may introduce plenty of noises, yielding disturbing multimodal context. To alleviate this issue, we exploit DPT to selectively suppress disturbing edges which do not belong to the DPT structure. Concretely, we compute a tree mask $S \in \mathbb{R}^{T \times T}$ to restrict the adjacency matrix A as follows:

$$S_{ij} = \begin{cases} 1, \ i \in \mathcal{C}(j) \ or \ j \in \mathcal{C}(i) \\ \\ \alpha, \ otherwise, \end{cases} \tag{6}$$

where $i, j \in [1, T]$ are nodes in the parsing tree, $\mathcal{C}(j)$ denotes the children nodes set of node j, and α is a hyperparameter which is set as 0.1 in our paper. Then we multiply the adjacency matrix A with the tree mask S elementwisely to obtain a soft tree propagation route A_t to diffuse information on the graph by:

$$A_t = A \odot S, \tag{7}$$

where $\odot$ is elementwise multiplication. We then adopt one graph convolution layer [19] to propagate and update node features as follows:

$$Z = (A_t + I)XW_z, \tag{8}$$

where I is an identity matrix serving as shortcut connection to ease optimization, $W_z \in \mathbb{R}^{C_h \times C_h}$ is the parameter for updating node features, and $Z \in \mathbb{R}^{T \times C_h}$ is the output of the graph convolution. After propagation, each word node can include valid multimodal context and exclude disturbing ones through the proper edges in parsing tree, forming robust features aligned with the whole sentence.

Distribute: Finally, we distribute the updated features of word graph nodes Z back to all the spatial locations using the transpose of B by:

$$\tilde{Z} = B^T Z. \tag{9}$$

We further conduct max pooling over word features $Q \in \mathbb{R}^{T \times C_l}$ to obtain sentence feature $L \in \mathbb{R}^{C_l}$, and then tile L for $H \times W$ times to form grid-like sentence feature $\hat{L} \in \mathbb{R}^{H \times W \times C_l}$. As shown in Fig. 2, the distributed feature $\tilde{Z} \in \mathbb{R}^{H \times W \times C_h}$ is concatenated with V, $\hat{L}$ and P and then fed into a 1×1 convolution to get the output feature $Y \in \mathbb{R}^{H \times W \times C_o}$.

3.3 Dual-Path Multi-Level Feature Fusion

It has been shown that the integration of features at different levels can lead to significant performance improvement of referring image segmentation [4,21,39]. We therefore also extract 4 levels of visual features $\{V_2, V_3, V_4, V_5\}$ as the input of our LSCM module. Then we utilize convolutional LSTM [35] to fuse the output features of the LSCM module $\{Y_2, Y_3, Y_4, Y_5\}$. The fusion process is illustrated in Fig. 2. We propose a Dual-Path Multi-Level Fusion module which sequentially fuses the features from 4 levels through the bottom-up and top-down paths. The input sequence of ConvLSTM is $[Y_5, Y_4, Y_3, Y_2, Y_3, Y_4, Y_5]$. The first bottom-up path sequentially integrates low-level features, which is able to complement high-level features with spatial details to refine the local parts of the mask. However, high-level features, which are critical for the model to recognize and localize the overall contour of the referred entities, are gradually diluted when integrating more and more low-level features. Thus, the top-down fusion path which reuses Y_3, Y_4 and Y_5 after bottom-up path is adopted to supplement more semantic multimodal information. Our Dual-Path Multi-Level Fusion module serves as a role to enhance features with both high-level semantics and low-level details for better segmentation performance.

4 Experiments

4.1 Experimental Setting

Datasets: We conduct extensive experiments on four benchmarks including UNC [41], UNC+ [41], G-Ref [28] and ReferIt [17]. UNC and UNC+ [41] are both collected from MS COCO dataset [24]. The UNC dataset contains $19,994$ images with $142,209$ referring expressions for $50,000$ objects while the UNC+ dataset contains $19,992$ images with $141,564$ expressions for $49,856$ objects. UNC+ has no location words hence it is more challenging than UNC. G-Ref [28] is also built upon the MS COCO dataset [24]. It consists of $26,711$ images with $104,560$ referring expressions for $54,822$ objects. The expressions are of average length of 8.4 words which is much longer than that of the other three datasets (with average length less than 4). ReferIt [17] is composed of $19,894$ images with $130,525$ referring expressions for $96,654$ objects. It also contains stuff categories.

Implementation Details: Following previous works [21,39], we choose DeepLab-ResNet101 [6] pre-trained on Pascal VOC dataset [10] as our backbone CNN. $Res2$, $Res3$, $Res4$ and $Res5$ are adopted for multi-level feature fusion. Input image is resized to 320 × 320. The maximum length of each referring expression is set to 20. For feature dimensions, we set $C_v = C_l = C_h = 1000, C_o = 500$. $\alpha = 0.1$ in our final model. The network is trained using Adam optimizer [18] with an initial learning rate of $2.5e^{-4}$ and a weight decay of $5e^{-4}$. We apply a polynomial decay with power of 0.9 to the learning rate. CNN is fixed during training. We use batch size 1 and stop training after $700K$ iterations. GloVe word embeddings [30] pretrained on Common Crawl with $840B$ tokens are used to replace randomly initialized ones. For fair comparison with prior works, all the final segmentation results are refined by DenseCRF [20].

Evaluation Metrics: Following the setup of prior works [4,16,21,39], we adopt overall intersection-over-union ($Overall\ IoU$) and precision with different thresholds ($Pr@X$) as the evaluation metrics for our model. The $Overall\ IoU$ is calculated by dividing the total intersection area with the total union area, where both intersection area and union area are accumulated over all test samples. The $Pr@X$ measures the percentage of prediction masks whose IoU is higher than the threshold X, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.2 Comparison with State-of-the-Arts

Table 1 summarizes the comparison results in $Overall\ IoU$ between our method and previous state-of-the-art methods. As illustrated in Table 1, our method consistently outperforms both bottom-up and top-down state-of-the-art methods on four benchmark datasets.

For bottom-up methods, STEP [4] densely fuses 5 feature levels for 25 times and achieves notable performance gains over CMSA [39]. Our method outperforms STEP on all the splits using less times of multimodal feature fusion, which indicates that our LSCM can capture more valid mulitmodal context information to better align features between visual and linguistic modalities. Particularly,

Table 1. Comparison with state-of-the-art methods on four benchmarks using *Overall IoU* as metric. “n/a” denotes methods does not use the same split as others. “BU” and “TD” denote “Bottom-Up” and “Top-Down” respectively.

Type	Method	ReferIt test	UNC			UNC+			G-Ref val
			val	testA	testB	val	testA	testB	
TD	MAttNet [40]	-	56.51	62.37	51.70	46.67	52.39	40.08	n/a
	CAC [8]	-	58.90	61.77	53.81	-	-	-	44.32
	NMTree [26]	-	56.59	63.02	52.06	47.40	53.01	41.56	n/a
BU	LSTM-CNN [16]	48.03	-	-	-	-	-	-	28.14
	DMN [29]	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
	RMI [25]	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
	KWA [32]	59.09	-	-	-	-	-	-	36.92
	CMSA(vgg16) [39]	59.91	52.38	54.68	49.59	34.41	36.53	30.10	32.35
	ASGN [31]	60.31	50.46	51.20	49.27	38.41	39.79	35.97	41.36
	RRN [21]	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
	Ours(vgg16)	63.82	55.41	57.92	52.54	41.18	44.32	35.78	39.78
	CMSA [39]	63.80	58.32	60.61	55.09	43.76	47.60	37.89	39.98
	STEP [4]	64.13	60.04	63.46	57.97	48.19	52.33	40.41	46.40
	Ours	**66.57**	**61.47**	**64.99**	**59.55**	**49.34**	**53.12**	**43.50**	**48.05**

ReferIt is a challenging dataset on which pervious methods only achieve marginal improvements. CMSA and STEP outperform RRN [21] by 0.17% and 0.50% IoU respectively, while our method significantly boost the performance gain to 2.94%, which well demonstrates the effectiveness of our method. Moreover, on UNC+ dataset which has no location words, our method also achieves 3.09% over STEP on testB split, showing that our method can model richer multimodal context information with less input conditions. In addition, we reimplement CSMA using their released code and our method using VGG16 as backbone. Our VGG16-based method also yields better performance on all 4 datasets with margins of 3.24% on UNC, 7.79% on UNC+, 7.43% on G-Ref and 3.91% on ReferIt dataset, showing that our method can well adapt to different visual features.

For top-down methods, MAttNet [40] and CAC [8] first generate a set of object proposals and then predict the foreground mask within the selected proposal. The decoupling of detection and segmentation relies on Mask-RCNN which is pretrained on much more COCO images ($110K$) than bottom-up methods using only PASCAL-VOC images ($10K$) for pretraining. Therefore, comparing their performances with bottom-up methods may not be completely fair. However, our method still outperforms MAttNet and CAC with large margins, indicating the superiority of our method. In addition, on ReferIt dataset which contains sentences about stuff, our method achieves state-of-the-art performance while top-down methods may not be able to well handle them.

There are also many top-down works [14,36–38] in referring localization field which adopt graphs to conduct grounding. Their graphs are composed of

Table 2. Ablation studies on UNC val set. All models use the same backbone (DeepLab-ResNet101) and DenseCRF for postprocessing. *The statistics of RRN-CNN [21] are higher than those reported in the original paper which do not use DenseCRF. Row 8 and row 11 are the same models with different names.

	Method	*Pr@0.5*	*Pr@0.6*	*Pr@0.7*	*Pr@0.8*	*Pr@0.9*	*IoU* (%)
1	RRN-CNN [21]*	46.99	37.96	27.86	16.25	3.75	47.26
2	+LSCM	61.26	52.93	43.39	27.38	6.70	54.87
3	+LSCM, GloVe [30]	63.13	54.20	43.38	27.54	6.78	55.93
4	+LSCM, GloVe, Mutan [2]	**64.25**	**55.64**	**45.00**	**29.24**	**7.28**	**56.50**
5	Multi-Level-RRN-CNN [21]*	65.83	57.45	46.76	31.91	10.40	57.61
6	+LSCM	68.33	61.16	51.59	36.98	11.57	59.67
7	+LSCM, GloVe [30]	70.56	62.89	52.91	38.07	11.99	60.98
8	+LSCM, GloVe, Mutan [2]	**70.84**	**63.82**	**53.67**	**38.69**	**12.06**	**61.54**
9	+Concat Fusion	68.49	60.78	50.92	34.87	9.94	60.10
10	+Gated Fusion [39]	69.08	62.46	50.73	35.42	11.27	60.46
11	+Dual-Path Fusion (Ours)	**70.84**	**63.82**	**53.67**	**38.69**	**12.06**	**61.54**

region proposals which rely on detectors pretrained on COCO and/or other large datasets. However, our DPT-WG consists of referring words and uses DPT to suppress disturbing edges in WG. Then, features of WG are distributed back to highlight grid format features of the referent for bottom-up mask prediction. Thus, our method is also different from NMTree [26] in which neural modules are assembled to tree nodes to conduct progressive grounding (i.e., retrieval) based on region proposals.

4.3 Ablation Studies

We perform ablation studies on UNC val set to verify the effectiveness of our proposed LSCM module and the Dual-Path Fusion module for leveraging multi-level features. Experimental results are summarized in Table 2.

LSCM Module: We first explore the effectiveness of our proposed LSCM module based on single level feature. Following [39], we implement the RRN-CNN [21] model without the recurrent refinement module as our baseline. Our baseline uses an LSTM to encode the whole referring expression as a sentence feature vector, and then concatenates it with each spatial location of the $Res5$ feature from DeepLab-101. Fusion and prediction are conducted on the concatenated features for generating final mask results. As shown in rows 1 to 4 of Table 2, **+LSCM** indicates that introducing our LSCM module into the baseline model can bring a significant performance gain of 7.61% IoU, demonstrating that our LSCM can well model valid multimodal context under the guidance of linguistic structure. Row 3 and Row 4 show that incorporating GloVe [30] and Mutan [2] fusion can further boost the performance based on our LSCM module.

We further conduct the same ablation studies based on multi-level features. All the models use our proposed Dual-Path Fusion module to fuse multi-level features. As shown in rows 5 to 8 of Table 2, our multi-level models achieve consistent performance improvements as the single-level models. These results well prove that our LSCM module can effectively capture multi-level context as well. Moreover, we additionally adapt **GloRe** [9] for the referring segmentation task over multi-level features and achieve 58.53% IoU and 67.23% *Pr*@0.5. The adapted GloRe uses learned projection matrix to project multimodal features into fixed number of abstract graph nodes, then conducts graph convolutions and reprojection to refine multimodal features. Our **+LSCM** in row 6 outperforms GloRe by 1.14% IoU and 1.10% *Pr*@0.5, indicating that building word graph by cross-modal attention and incorporating DPT to suppress disturbing edges between word nodes can better model valid multimodal context than GloRe.

Multi-level Feature Fusion: We compare different methods including Concat Fusion, Gated Fusion [39] and our Dual-Path Fusion for multi-level feature fusion. All the fusion methods take 4 levels of multimodal features processed by our LSCM module as input. As shown in rows 9 to 11 of Table 2, our proposed Dual-Path Multi-Level Fusion module achieves the best result, showing the effectiveness of integrating both high-level semantics and low-level details. In addition, the gated fusion from [39] conducts 9 fusion operations while ours conducts 6 fusion operations with better performance.

Table 3. Experiments of graph convolution in terms of *Overall IoU*. n denotes number of layers of graph convolution in our LSCM module. $\alpha = 0.1$ here.

	+LSCM, GloVe					
	$n = 0$	$n = 1$	$n = 2$	$n = 3$	$n = 4$	adaptive
UNC val	54.51	**55.93**	50.77	50.64	49.59	54.69
G-Ref val	38.94	**40.54**	39.29	37.74	37.50	37.41

Layers of Graph Convolution: In Table 3, we explore the effects of conducting different layers of graph convolution in our LSCM module on UNC val set and G-Ref val set. The results show that the naive increase of graph convolution layers in LSCM will deteriorate the segmentation performance, probably because multiple rounds of message propagation among all words muddle the multimodal context of each word instead of enhancing it. Besides, adaptive which means number of the graph convolution layers equal to the depth of DPT, yields lower performance than one layer of graph convolution. It indicates that propagating information among word nodes without further constrain will include more disturbing context. Conducting 1 layer of graph convolution to communication between parents and children nodes is already sufficient without introducing too much noises, which also makes our method more efficient. In addition, $n = 1$ also outperforms $n = 0$ which shows communication among words is necessary after gathering multimodal context for each word.

Edge Weights in Tree Mask: In Table 4, we explore how different values of α in the tree mask S (Eq. 6) influence the performance of our LSCM. We can

Table 4. *Overall IoU* results of different edge weights α in tree mask S. Experiments are conducted on UNC val set. All the models use $n = 1$ layer of graph convolution.

+LSCM, GloVe						
$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 1$
55.01	**55.93**	55.44	55.49	54.77	55.12	54.81

observe that $\alpha = 0.1$ achieves the best performance and outperforms $\alpha = 1$ (WG w/o DPT) by 1.12% IoU, which demonstrates that suppressing syntactic irrelevant edges in our word graph can reduce unnecessary information propagation and exclude disturbing multimodal context. In addition, $\alpha = 0$ yields inferior performance to $\alpha = 0.1$, indicating our DPT-WG (i.e., approximate spanning tree) can obtain more sufficient information than a strict DPT.

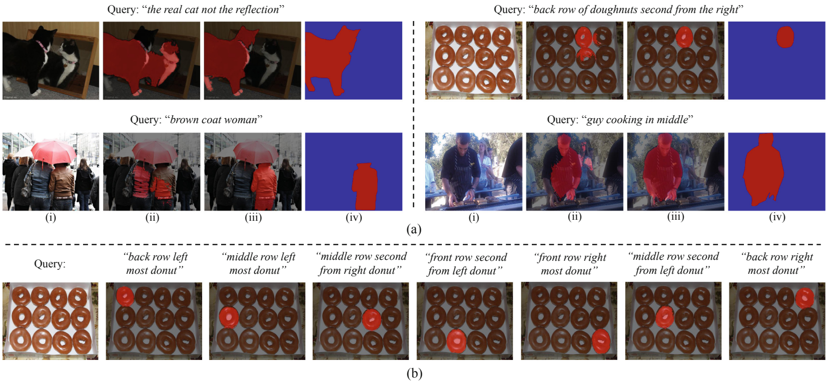
Query: "the real cat not the reflection"
Query: "back row of doughnuts second from the right"
Query: "brown coat woman"
Query: "guy cooking in middle"
(i)
(ii)
(iii)
(iv)
(i)
(ii)
(iii)
(iv)
(a)
Query:
"back row left most donut"
"middle row left most donut"
"middle row second from right donut"
"front row second from left donut"
"front row right most donut"
"middle row second from left donut"
"back row right most donut"
(b)

Fig. 4. Qualitative Results of referring image segmentation. (a)(i) Original image. (a)(ii) Results produced by the multi-level RRN-CNN baseline (row 5 in Table 2). (a)(iii) Results produced by our full model (row 8 in Table 2). (a)(iv) Ground-truth. (b) Results of customized expressions. Our model can adapt to new expressions flexibly.

Qualitative Results: Figure 4(a) presents the segmentation results predicted by our full model (row 8 in Table 2) and the multi-level RRN-CNN baseline (row 5 in Table 2). Comparing (b) and (c) in Fig. 4, we can find that only multi-level feature refinement without valid multimodal context modeling is not sufficient for the model to understand the referring expression comprehensively, thus resulting in inaccurate predictions, such as segmenting "coat" but ignoring "brown" in the bottom-left of Fig. 4. As shown in Fig. 4(b), we also manually generate customized expressions to traverse many the donuts. It is interesting to find that our model can always understand different expressions adaptively and locate the right donuts, indicating that our model is flexible and controllable. More qualitative results on four datasets are presented in supplementary materials.

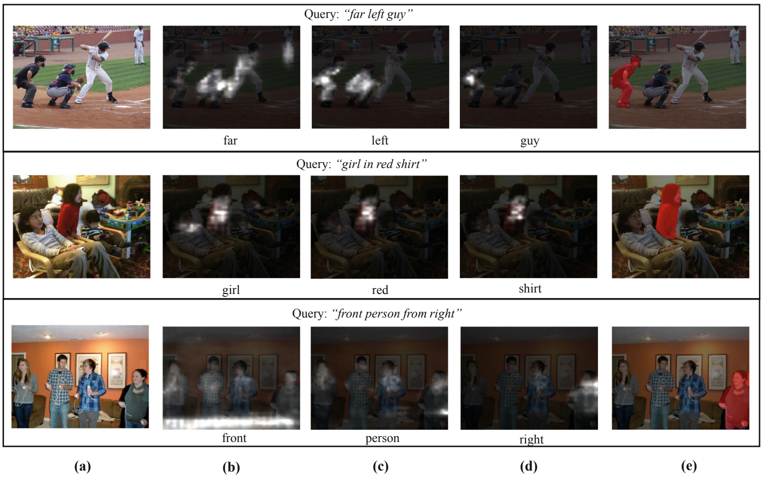

Fig. 5. Visualization of attention maps on the given words. (a) the original image. (b)(c)(d) refer to the attention maps of the specific words below. (e) predictions of our proposed method.

Visualization of Attention Maps: To give a straightforward explanation about how our LSCM works, we visualize the attention maps of each node (corresponding to the words of referring expression) to the spatial locations and the results are shown in Fig. 5. The cross-modal attention maps correspond to B obtained in the gather operation (Eqs. 1 and 2), which has size of $T \times HW$. Each row of B denotes the attention map of a certain word. The three words are organized in sequential order. From Fig. 5 we find that a meaningful word usually attends to its corresponding area in the image. For example, in the third row of (b), the word "front" attends to the front area of the image, and in the second row of (c), word "red" attends to the area of red shirt. Our LSCM module is able to model valid multimodal context among these attended areas to obtain a precise segmentation of the referring expression.

5 Conclusion and Future Work

In this paper, we explore the referring image segmentation problem by introducing a "gather-propagate-distribute" scheme to model multimodal context. We implement this scheme as a Linguistic Structure guided Context Modeling (LSCM) module. Our LSCM builds a Dependency Parsing Tree suppressed Word Graph (DPT-WG) which guides all the words to include valid multimodal context of the sentence while excluding disturbing ones, which can effectively highlight multimodal features of the referent. Our proposed model achieves state-of-the-art performance on four benchmarks. In the future, we plan to adapt our LSCM module into other tasks (e.g., VQA, Captioning) to verify its effectiveness.

Acknowledgement. This work is supported by Guangdong Basic and Applied Basic Research Foundation (No. 2020B1515020048), National Natural Science Foundation of China (Grant 61876177, Grant 61976250), Beijing Natural Science Foundation (L182013, 4202034), Fundamental Research Funds for the Central Universities, Zhejiang Lab (No. 2019KD0AB04) and Tencent Open Fund.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. TPAMI (2017)
2. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: multimodal tucker fusion for visual question answering. In: ICCV (2017)
3. Chen, D., Manning, C.D.: A fast and accurate dependency parser using neural networks. In: EMNLP (2014)
4. Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: ICCV (2019)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **40**(4), 834–848 (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, Y.W., Tsai, Y.H., Wang, T., Lin, Y.Y., Yang, M.H.: Referring expression object segmentation with caption-aware consistency. arXiv preprint arXiv:1910.04748 (2019)
9. Chen, Y., et al.: Graph-based global reasoning networks. In: CVPR (2019)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. IJCV **88**, 303–338 (2010)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
14. Hu, R., Rohrbach, A., Darrell, T., Saenko, K.: Language-conditioned graph networks for relational reasoning. In: ICCV (2019)
15. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: CVPR (2017)
16. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9905, pp. 108–128. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_7
17. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferitGame: referring to objects in photographs of natural scenes. In: EMNLP (2014)
18. Kingma, D.P., Ba, J.: Adam: amethod for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
20. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: NeurIPS (2011)
21. Li, R., et al.: Referring image segmentation via recurrent refinement networks. In: CVPR (2018)
22. Liao, Y., et al.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: CVPR (2020)
23. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
24. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, C., et al.: Recurrent multimodal interaction for referring image segmentation. In: ICCV (2017)
26. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: ICCV (2019)
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
28. Mao, J., et al.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
29. Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11215, pp. 656–672. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_39
30. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: EMNLP (2014)
31. Qiu, S., Zhao, Y., Jiao, J., Wei, Y., Wei, S.: Referring image segmentation by generative adversarial learning. IEEE Trans. Multimedia (TMM) **22**(5), 1333–1344 (2019)
32. Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11210, pp. 38–54. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_3
33. Vaswani, A., et al.: Attention is all you need. In: NeurIPS (2017)
34. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
35. Xingjian, S., et al.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: NeurIPS (2015)
36. Yang, S., Li, G., Yu, Y.: Cross-modal relationship inference for grounding referring expressions. In: CVPR (2019)
37. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: ICCV (2019)
38. Yang, S., Li, G., Yu, Y.: Graph-structured referring expression reasoning in the wild. In: CVPR (2020)
39. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR (2019)
40. Yu, L., et al.: MAttNet: modular attention network for referring expression comprehension. In: CVPR (2018)

41. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9906, pp. 69–85. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_5
42. Zhang, H., et al.: Context encoding for semantic segmentation. In: CVPR (2018)
43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)