Locate then Segment: A Strong Pipeline for Referring Image Segmentation

Ya Jing,[1,2*] Tao Kong,[3] Wei Wang,[1,2] Liang Wang,[1,2] Lei Li,[3] Tieniu Tan[1,2]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3]ByteDance AI Lab

Abstract

Referring image segmentation aims to segment the objects referred by a natural language expression. Previous methods usually focus on designing an implicit and recurrent feature interaction mechanism to fuse the visual-linguistic features to directly generate the final segmentation mask without explicitly modeling the localization information of the referent instances. To tackle these problems, we view this task from another perspective by decoupling it into a "Locate-Then-Segment" (LTS) scheme. Given a language expression, people generally first perform attention to the corresponding target image regions, then generate a fine segmentation mask about the object based on its context. The LTS first extracts and fuses both visual and textual features to get a cross-modal representation, then applies a cross-model interaction on the visual-textual features to locate the referred object with position prior, and finally generates the segmentation result with a light-weight segmentation network. Our LTS is simple but surprisingly effective. On three popular benchmark datasets, the LTS outperforms all the previous state-of-the-arts methods by a large margin (e.g., +3.2% on RefCOCO+ and +3.4% on RefCOCOg). In addition, our model is more interpretable with explicitly locating the object, which is also proved by visualization experiments. We believe this framework is promising to serve as a strong baseline for referring image segmentation.

1. Introduction

Jointly learning vision and language is a significant task in machine learning and pattern recognition community, which has drawn great attention in recent years. In this paper, we study the challenging task of language-instructed object segmentation [11, 38, 37] which aims to generate a segmentation mask of the object in image referred by a natural language expression. It has wide applications, e.g., interactive image editing and language-guided human-robot interaction. Beyond traditional semantic segmentation, language-referring image segmentation is more challenging due to the semantic gap between image and language. In addition, the textual expression is not just limited to entities (e.g., "person", "horse"). It may contain descriptive words, such as object properties (e.g., "red", "young"), actions (e.g., "standing", "hold"), and positional relationships (e.g., "right", "above").

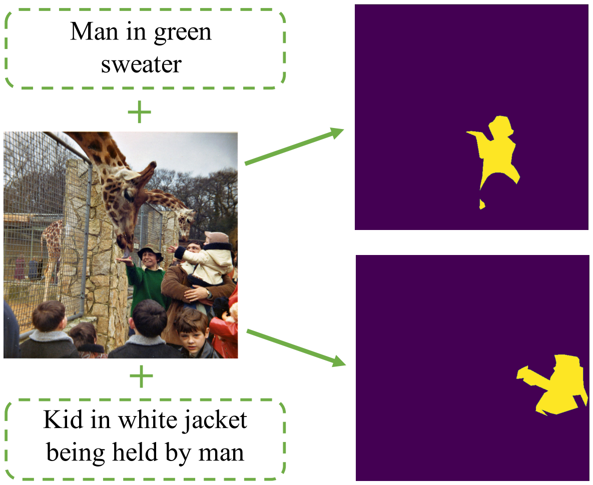


Figure 1. The illustration of referring image segmentation. Given a referring expression and an image, the model aims to generate a segmentation mask of the corresponding object in image referred by the language expression. Best viewed in color.

Given the image and referring sentence, there are two essential issues affecting the overall performance of a referring image segmentation model. First, the model must highlight the most discriminative candidate area in image corresponding to the given language. Second, the model must generate a fine segmentation result. The existing referring image segmentation methods could be generally sum-

[*] This work was done when Ya Jing was an intern at ByteDance AI Lab.

marized as follows: (1) Utilizing a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to represent the image feature $f_v(I)$ and language feature $f_{text}(X)$, respectively. (2) Cross-modal attention and recurrent ConvLSTM are used to fuse $f_v(I)$ and $f_{text}(X)$ to get a coarse mask. (3) Dense CRF (DCRF) is further used as post-processing to get the final fine segmentation $M(I)$.

Previous works mainly focus on how to fuse the image feature and language feature. A straightforward solution [11] is to utilize a concatenation-and-convolution method to fuse visual and linguistic representations to produce the final segmentation result. However, this method cannot model the alignment between image and language effectively due to the fact that the visual and textual information is modeled individually. To further model the context between multi-modal features, some prior methods [32, 4, 37] propose cross-modal attention by adaptively focusing on important regions in the image and informative keywords in the language expression. Recently, to exploit different types of informative keywords in the language and learn the aligned multi-modal representations, some works [13, 14] either perceive all the entities that are referred by the expression or utilize the linguistic structure as guidance to segment the referent. Although great progress has been made, the network architecture and experimental practice have steadily become more and more complex. This makes the algorithm analysis and comparison more and more difficult. In addition, they do not explicitly locate the referred object guided by language expression and only utilize time-consuming post-processing DCRF to generate the final refined segmentation.

In this paper, we consider solving this problem from another perspective. We decouple the referring image segmentation task into two sub-sequential tasks: (a) referring object position prediction, and (b) object segmentation mask generation. In our model, we first fuse the visual and linguistic features to get a cross-modal feature. Then for (a), we propose a localization module to directly obtain the visual contents prior corresponding to the expression. Such object prior will be used as a visual positional guidance for the subsequent segmentation module. For (b), we concatenate the object prior with the cross-modal features and utilize a light-weight ConvNets to get the final segmentation mask (Fig. 2).

Our solution is very simple but surprisingly effective. On three challenging benchmarks, i.e., RefCOCO [15], RefCOCO+ [15] and RefCOCOg [27], our model outperforms the state-of-the-arts methods by a large margin (*e.g.,* +3.2% on RefCOCO+ and +3.4% on RefCOCOg). Extensive ablation studies also verify the effectiveness of each component of our method.

2. Related Work

In this section, we briefly review the related work about prior studies on object segmentation, referring image localization and segmentation, and cross-model interaction.

2.1. Object Segmentation

Object segmentation has achieved great advances in recent years based on Fully Convolutional Network (FCN) [24]. FCN-based models transform fully connected layers in CNN into convolutional layers to train a segmentation model in an end-to-end way. DeepLab [5] replaces regular convolution with atrous (dilated) convolution to enlarge the receptive field of filters, leading to larger feature maps with richer semantic information. PSPNet [39] proposes a pyramid pooling module to capture multi-scale information. Some other works [1, 20] exploit low level features containing detailed information to generate more accurate results. The instance segmentation area also achieved great progress based on Mask R-CNN [8] and FCNs [24, 35]. In this paper, we study the more challenging segmentation problem whose semantic categories are referred by language expression.

2.2. Referring Localization and Segmentation

Referring image localization aims to localize specific objects in an image referred by a language expression with a bounding box. Some works [36, 10] model the relationships between image and language to obtain the most related objects. MAttNet [38] decomposes the referring expression into subject, location and relationship to compute a more accurate matching score. The aim of referring image segmentation[11] is to localize the referred object with a segmentation mask rather than a bounding box. Hu et al. [11] utilize the concatenation of visual and linguistic features from CNN and Long Short-Term Memory network (LSTM) [9] to generate the segmentation mask. To obtain more accurate result, [19] fuses multi-level visual features to refine the local details of segmentation mask. Multi-modal LSTM [22] is employed to sequentially fuse visual and linguistic features in multiple time steps. Dynamic filters [28] for each word further enhance multi-modal features. Shi et al. [32] utilize word attention to model keyword-aware context. Recently, some works [13, 14] either perceive all the entities that are referred by the expression or utilize the linguistic structure as guidance to segment the referent. Multi-task collaborative network [26] achieves a joint learning of referring expression comprehension and segmentation. In this paper, we propose a localization module to locate the referred object with position prior and a segmentation module to obtain the final segmentation result.

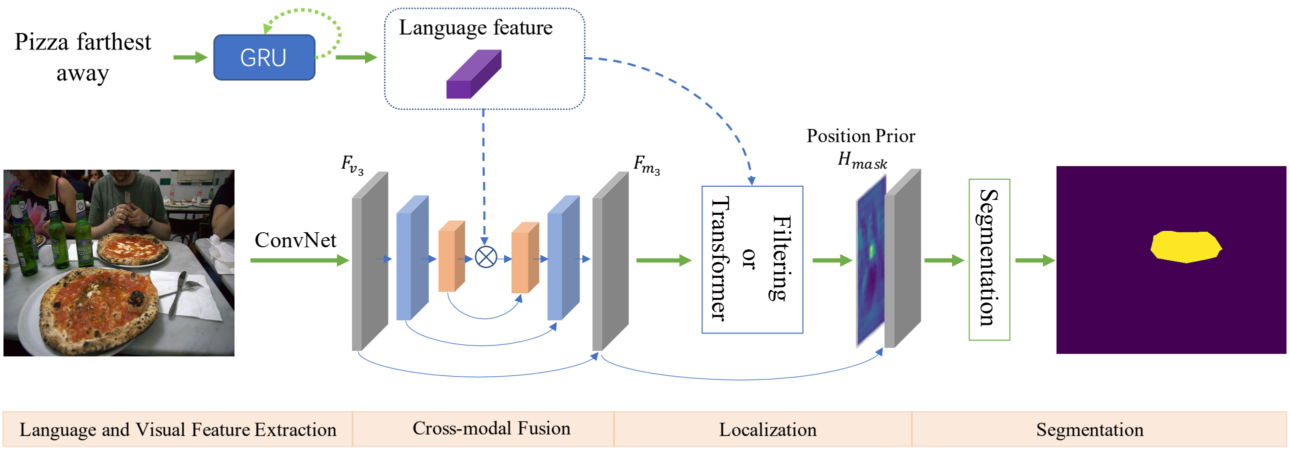


Figure 2. The architecture of our proposed method. The visual feature and linguistic feature are extracted by a deep convolutional network (ConvNet) and a bi-GRU network respectively, and then fused to generate the cross-modal features. Next a cross-modal interaction module (e.g., filtering and transformer [33]) is proposed to generate the object position prior. Finally, we concatenate the position prior and the cross-modal features to generate final segmentation mask by further convolutional refinement.

2.3. Cross-model Interaction

In cross-modal tasks, a main challenge is to model the relationship between image and text. Recently, attention mechanism has been shown to be a powerful technique to extract the visual contents corresponding to the language expression in referring image segmentation.

The relevance filtering can be seen as a simple way of attention mechanism, which is widely used in different areas of computer vision. Object tracking [3] aims to localize an object in a video given the object region in the first frame, where relevance filtering is used to compare the first frame with the rest ones. Object classification [34] can be seen as a relevance filtering produce between output image feature and weight matrix of the last layer. Previously, relevance filtering has been considered in referring image segmentation [28], but they use it implicitly to generate the final segmentation mask. In this paper, we utilize the direct language-conditional relevance filtering to obtain the relevance heatmap where higher response value is directly considered as the referred object prior.

In addition to filtering, many cross-modal attention models [32, 4, 37] are proposed to adaptively focus on important regions in the image and informative keywords in the language expression. Different from them, we propose to utilize the unified attention-based building block transformer [33] to get the cross-modal relevance, which eliminates the need to design complex attention models.

3. Proposed Approach

In this section, we explain the proposed LTS in detail. First, we introduce the procedure of visual and textual representations extraction. Then we describe the two modules including filtering (or transformer) based localization and light-weight ConvNets based segmentation. Finally, we give the details of model learning process.

3.1. Visual and Linguistic Feature Extraction

As shown in Fig. 2, the input of our model consists of an image I and a referring expression X. For similicity, we utilize ConvNets and GRUs to extract features of I and X respectively.

Visual Feature For the input image $I \in \mathbb{R}^{H\times W\times 3}$, we utilize the visual backbone to extract the multi-level visual features, which are denoted as $F_{v_1} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times d_1}$, $F_{v_2} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times d_2}$, and $F_{v_3} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times d_3}$, respectively. Note that d is the dimension of feature channel, H and W are the height and width of the original image, respectively.

Linguistic Feature Given a referring sentence $X = [x_1, x_2, ..., x_m]$, where x_i is the i-th token. We first apply table lookup to obtain the word embeddings. The embeddings are initialized as a 300-dimensional embedding vector by GLOVE embeddings [30]. To model the dependencies between adjacent words, we use the standard bi-directional Gated Recurrent Unit (GRU) [7] to handle the initial embedding textual vectors:

$$\overrightarrow{h_t} = \overrightarrow{GRU}(x_t, \overrightarrow{h_{t-1}}), h_0 = 0, \tag{1}$$

$$\overleftarrow{h_t} = \overleftarrow{GRU}(x_t, \overleftarrow{h_{t+1}}), h_{m+1} = 0, \tag{2}$$

where $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ represent the forward and backward GRUs, respectively. The global textual representation is obtained by average pooling between all word representations, which is donated as:

$$f_{text} = avg(h_1, h_2, ..., h_m), \tag{3}$$

$$h_t = concat(\overrightarrow{h_t}, \overleftarrow{h_t}), t \in [1, 2, ..., m], \tag{4}$$

Fusion We obtain the multi-modal tensor by fusing F_{v_1} with f_{text}, which is formulated as:

$$f^l_{m_1} = g(f^l_{v_1} W_{v_1}) \cdot g(f_{text} W_t), \tag{5}$$

where g denotes Leaky ReLU, $f^l_{m_1}$ and $f^l_{v_1}$ are the feature vectors of F_{m_1} and F_{v_1}, respectively, W_{v_1} and W_t are two transformation matrices to transform the visual and textual representations into the same feature dimension. Then, the multimodal tensors, F_{m_2} and F_{m_3} are obtained by:

$$F^{'}_{m_{i-1}} = UpSample(F_{m_{i-1}}) \tag{6}$$

$$F_{m_i} = concat(g(F^{'}_{m_{i-1}} W_{m_{i-1}}), g(F_{v_i} W_{v_i})), \tag{7}$$

where $i \in [2, 3]$ and UpSampling has a stride of 2 × 2. In the following process, we utilize F_{m_3} as the input to generate the segmentation mask. Previous works usually adopt recurrent attention mechanism to get the segmentation results. In this paper, we show that locate-then-segment could get surprisingly superior performance, which will be introduces as follows.

3.2. Localization

To locate the object referred by language expression, we propose two ways to capture the context between multimodal features including simple relevance filtering and unified attention-based block transformer, which eliminates the need to design complex attention model.

Relevance Filtering The feature F_{m_3} contains rich cross-modal information, which must be further modeled to get the relevant area in the image. The aim of our cross-modality relevance filtering is to find the visual regions referred by the language expression, whose response scores are higher than the unrelated regions. We first generate the language-guided kernel $K = f_{text} W_k$, where $K \in \mathbb{R}^{d_k}$. Then it is reshaped into $\mathbb{R}^{d_k \times 1 \times 1}$ to perform filtering in fusion feature F_{m_3}:

$$H_{mask} = conv(K, F_{m_3}), \tag{8}$$

where $H_{mask} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8}}$ and $conv$ means convolution operation. The heatmap H_{mask} is a coarse segmentation mask where regions with higher response score means the more likely corresponding to the language expression (see Fig. 2 position prior).

Transformer To maintain consistency with the relevance filtering, here we do not utilize the transformer encoder to extract the textual representations but regard the global textual representation f_{text} as the encoder output.

The decoder follows the standard architecture of the transformer, transforming the multimodal feature F_{m_3} to response map H_{mask} using multi-headed attention mechanisms:

$$H_{mask} = decoder(F_{m_3}, f_{text}). \tag{9}$$

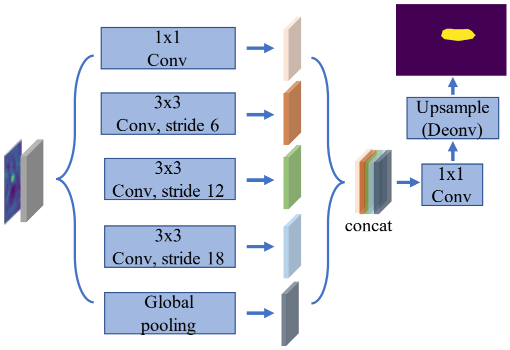


Figure 3. The segmentation module. We first concatenate the feature F_{m_3} and position prior map, and feed them into a single ASPP module, finally we upsample (deconvolution) the final generated mask.

The decoder expects a sequence as input, hence we collapse the spatial dimensions of F_{m_3} into one dimension, resulting in a $d \times \frac{HW}{64}$ feature map. Since the transformer architecture is permutation-invariant, we supplement it with fixed positional encodings [2] that are added to the input of each attention layer.

3.3. Segmentation

Given the visual object prior generated by Eq. (8) or Eq. (9), the aim of the segmentation module is to generate the final fine segmentation mask.

We first concatenate the original cross-modal feature F_{m_3} and visual object prior H_{mask}, and utilize a segmentation module to refine the coarse segmentation result:

$$P_{mask} = Seg(concat(F_{m_3}, H_{mask})), \tag{10}$$

where the main structure of Seg is ASPP [5]. The ASPP probes an incoming convolutional feature layer with filters at multiple sampling rates and effective fields-of-views, thus capturing objects as well as image context at multiple scales. Note that to obtain more precise segmentation results, we adopt the deconvolution method to upsample the feature map by a factor 2. Therefore, the predicted mask $P_{mask} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4}}$. Fig. 3 shows the segmentation process.

3.4. Training and Inference

During training, the Sigmoid Binary Cross Entropy (BCE) loss function is defined as follows:

$$L_{seg} = \sum_{l=1}^{\frac{H}{4} \times \frac{W}{4}} [y_l \log(p_l) + (1 - y_l) \log(1 - p_l)], \tag{11}$$

where y_l and p_l are the elements of the down-sampled ground-truth mask and predicted mask P_{mask}, respectively.

In addition, to make sure that the model can focus on the corresponding image regions, we add a locating loss to supervise the position prediction, which is defined as follows:

$$L_{loc} = \sum_{l=1}^{\frac{H}{8}\times\frac{W}{8}} [y_l \log(h_l) + (1-y_l)\log(1-h_l)], \quad (12)$$

where h_l is the element of the down-sampled response map H_{mask}.

Finally, the total loss is defined as:

$$L = L_{seg} + \lambda L_{loc}, \quad (13)$$

where λ is empirically set to 0.1 in our experiments.

During inference, we upsample the predicted segmentation mask P_{mask} to the original image size $H \times W$ and binarized at a threshold of 0.25 as the final result. No other post processing operations are needed.

4. Experiments

In this section, we first introduce the experimental setup including dataset, evaluation metrics, and implementation details. Then, we analyze the quantitative results of our method and a set of baseline variants. Finally, we visualize several segmentation masks.

4.1. Experimental Setup

Datasets and Metrics We evaluate the proposed method on three benchmark datasets, i.e., RefCOCO [15], RefCOCO+ [15] and RefCOCOg [27]. We adopt intersection-over-union (IoU) and prec@X as the evaluation metrics [26, 37]. The IoU calculates intersection regions over union regions of the predicted segmentation mask and the ground truth. The prec@X measures the percentage of test images with an IoU score higher than the threshold γ , where $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

The RefCOCO dataset contains 19,994 images with 142,210 referring expressions for 50,000 objects. The images and expressions are collected from the MSCOCO [21] with a two-player game [15]. It is split into train, validation, test A and test B with a number of 120,624, 10,834, 5,657 and 5,095 samples, respectively. In general, each image contains two or more objects with the same object class and each expression has an average length of 3.5 words.

The RefCOCO+ dataset contains 19,992 images with 141,564 referring expressions for 49,856 objects. The images and expressions are also collected from the MSCOCO. It is also split into train, validation, test A and test B with a number of 120,191, 10,758, 5,726 and 4,889 samples, respectively. Different from RefCOCO, the expressions in RefCOCO+ include more appearances than absolute locations.

The RefCOCOg dataset is also collected from MSCOCO and contains 26,711 images with 104,560 referring expressions for 54,822 objects. Different from RefCOCO and RefCOCO+, expressions in RefCOCOg are collected from Amazon Mechanical Turk instead of a two-player game and have a longer length of 8.4 words includes both appearances and locations of the referent. RefCOCOg [27, 29] have two types of data partitions, i.e., google partition [27] and UNC partition [29]. We adopt UNC partition in this paper.

Implementation Details Following previous work [26], we adopt the Darknet53 [31] as the visual backbone, which is pre-trained on MSCOCO while removing the images appeared in the validation and test sets of three datasets. The input images are resized to 416×416 and the input sentences are set with a maximum sentence length of 15 for RefCOCO and RefCOCO+, and 20 for RefCOCOg. A 1024 dimensional bi-GRU is used to extract the textual feature. The filtering dimension d_k is set to 1024. The decoder has 1 layer network, 4 heads and 1024 hidden units. Adam [16] is used as the optimizer to train our model. The initial learning rate is 0.001, which is decreased by a factor of 0.1 at 30-th epoch. The batch size and training epochs are set to 18 and 45, respectively.

4.2. Main Results

To demonstrate the effectiveness of our model, we compare our segmentation results with the state-of-the-arts (SOTAs) methods on three referring segmentation benchmarks utilizing relevance filtering as the localization module, as shown in Tab. 1. It can be seen that our model achieves the best performances under IoU metric across different datasets even though we do not utilize the time-consuming post-processing, e.g., DenseCRF [18] and ASNLS [26]. Note that our model can further improve the performance when adopting relevance filtering for two times as shown in Tab. 5. Specifically, compared with the best competitor CGAN [25] which proposes a cascade grouped attention network to perform step-wise reasoning, our model significantly outperforms it by about 3% absolute IoU point on two challenging datasets (RefCOCO+ and RefCOCOg) performing only the simple relevance filtering once. The improved performances over the best competitor indicate that our model is very effective for this task.

Compared with the methods CMPC [13] and LSCM [14] which either perceive all the entities that are referred by the expression or utilize the linguistic structure as guidance to segment the referred object, our model achieves much better performances by explicitly modeling the object position prior followed with a segmentation module, demonstrating the effectiveness of our pipeline. Here DMN [28] and Lang2seg [6] also model the multi-modal context by filtering. But DMN utilizes every word to generate the kernel and performs convolution. Therefore, it needs to regress all the generated maps. Lang2seg utilizes the spatial-aware dynamic filters and needs to perform filtering in 7 times for different local regions. Different from them, our LTS uti-

Table 1. Comparison with the state-of-the-art methods on three benchmarks datasets using IoU as metric. "-" represents that the result is not provided. DCRF and ASNLS means DenseCRF [18] and ASNLS [26] post-processings, respectively.

Methods	Backbone	DCRF	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
RMI [22]	ResNet101	✓	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [28]	ResNet101	✗	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN [19]	ResNet101	✓	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [38]	MRCN-ResNet101	✗	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [23]	MRCN-ResNet101	✗	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA [37]	ResNet101	✓	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2seg [6]	ResNet101	✗	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAM [12]	ResNet101	✓	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC [13]	ResNet101	✓	61.36	64.53	59.64	49.56	53.44	43.23	-	-
MCN+ASNLS [26]	DarkNet53	✗	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
LSCM [14]	ResNet101	✓	61.47	64.99	59.55	49.34	53.12	43.50	-	-
CGAN [25]	DarkNet53	✗	64.86	**68.04**	62.07	51.03	55.51	44.06	51.01	51.69
LTS (Ours)	DarkNet53	✗	**65.43**	67.76	**63.08**	**54.21**	**58.32**	**48.02**	**54.40**	**54.25**

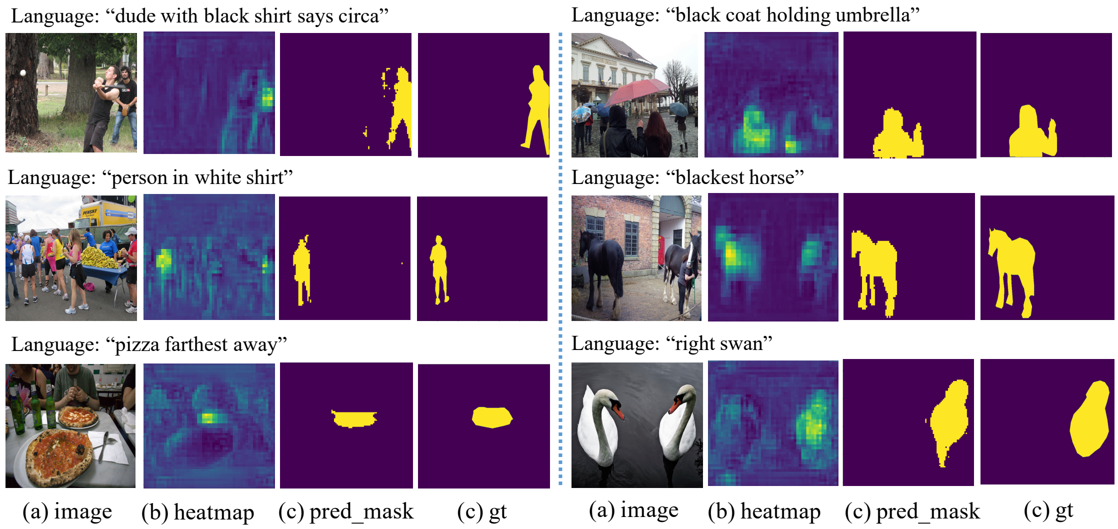

Language: "dude with black shirt says circa"
Language: "black coat holding umbrella"
Language: "person in white shirt"
Language: "blackest horse"
Language: "pizza farthest away"
Language: "right swan"
(a) image
(b) heatmap
(c) pred_mask
(c) gt
(a) image
(b) heatmap
(c) pred_mask
(c) gt

Figure 4. Visualization of correlation heatmaps H_{mask} (generated by relevance filtering) and final results P_{mask} (pred_mask) predicted by our model. gt means the ground truth segmentation mask of input image. Best viewed in color.

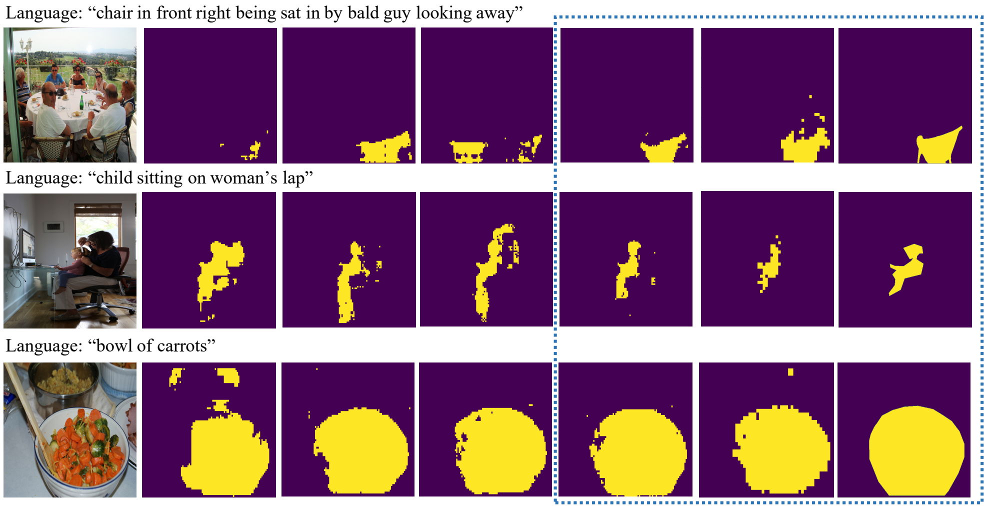
lizes the sentence feature to generate one kernel and only needs filtering once. And the heatmap is utilized as the location to guide the generation of segmentation mask. The improved performance (16% and 6% IoU point) over them suggests that our segmentation module can learn a more accurate segmentation mask after obtaining the object location by language expression.

4.3. Qualitative Results

To verify whether the proposed localization module can obtain the correct location of the referent and how the segmentation module works, we visualize the response heatmap H_{mask} (generated by relevance filtering) and segmentation results of several samples shown in Fig. 4. We can see that our model is able to localize the referent by the language-dependent filtering. We also evaluate the localization quality (the probability that the maximum value of heatmap lies in the ground truth mask). The result is 81.74%, which further demonstrates our localization capacity of the relevance filtering. However, this heatmap is far from an accurate segmentation mask. After refined by the segmentation module which captures objects and image context at multiple scales, we can obtain the final precise prediction mask.

Fig. 5 shows the segmentation masks obtained by different models, which demonstrates the benefits of each module in our proposed method.

- The predicted mask from model w/o segmentation can only obtain the location of the referent but not the fine mask of object, as shown in column (b);
- The model w/o fusion or filter also generates lower


Language: “chair in front right being sat in by bald guy looking away”
Language: “child sitting on woman’s lap”
Language: “bowl of carrots”

(a) image (b) w/o segmentation (c) w/o fusion (d) w/o filter (e) ours (f) MCN (g) gt

Figure 5. Qualitative examples of referring image segmentation by different models. (b) (c) (d) show the proposed model w/o segmentation, fusion, and filter, respectively. MCN is the method proposed in [26]. Best viewed in color.

Table 2. Ablation studies on RefCOCO dataset. Seg means the segmentation module.

	Fusion	Filter	Seg	L_{loc}	IoU	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9
val	✓				55.08	60.58	51.26	41.55	27.15	6.76
		✓			54.53	60.43	50.00	39.03	24.95	5.85
	✓	✓			56.94	63.49	54.20	44.13	29.43	8.36
	✓		✓		63.50	72.84	65.85	57.61	42.33	13.41
		✓	✓		63.64	72.94	66.52	58.01	42.84	13.20
	✓	✓	✓		65.05	75.01	68.48	60.69	44.93	14.03
	✓	✓	✓	✓	**65.43**	**75.16**	**69.51**	**60.74**	**45.17**	**14.41**
testA	✓				56.82	62.49	53.60	42.87	28.27	6.65
		✓			56.05	61.66	51.76	40.32	25.61	5.67
	✓	✓			58.53	64.52	55.65	45.59	30.51	7.94
	✓		✓		65.31	75.32	69.45	60.46	44.76	11.38
		✓	✓		66.41	77.00	71.27	62.63	46.65	12.82
	✓	✓	✓		67.49	78.10	72.87	64.47	**48.31**	**13.24**
	✓	✓	✓	✓	**67.76**	**78.47**	**73.13**	**64.56**	47.98	12.92
testB	✓				53.52	59.20	49.64	39.20	26.22	8.58
		✓			52.55	56.45	46.91	37.45	24.10	8.03
	✓	✓			55.12	60.61	51.21	41.32	29.07	10.64
	✓		✓		60.97	69.13	62.24	53.39	39.65	15.68
		✓	✓		60.72	68.03	61.14	52.60	39.59	16.68
	✓	✓	✓		62.43	70.99	**64.65**	55.27	42.41	**17.48**
	✓	✓	✓	✓	**63.08**	**71.82**	64.59	**55.74**	**42.79**	17.35

quality masks compared with our proposed full model as they fail to make clear judgement of the referred object, as shown in column (c) and (d).

Here we also compare our results with MCN [26] in column (e) and (f). Our generated segmentation masks have more obvious object shapes and finer outlines.

4.4. Ablation Experiments

Our proposed model is mainly composed of three modules, cross-modal fusion, localization (relevance filtering or transformer) and segmentation. To investigate these three components and the proposed locating loss in our model, we perform a set of ablation studies on the Refcoco dataset. Tab. 2 shows the result.

Table 3. Results of utilizing segmentation module on CMPC [13] on the RefCOCO dataset using IoU as metric.

Model	val	testA	testB
CMPC	61.36	64.53	59.64
CMPC+Seg	**62.75**	**65.34**	**61.08**

We first investigate the importance of fusing textual

Table 4. Results of utilizing transformer instead of filtering on the val sets of three datasets using IoU as metric.

Model	RefCOCO	RefCOCO+	RefCOCOg
LTS	65.43	54.21	54.40
LTS-Trans	**66.15**	**54.52**	**54.51**

Table 5. Effects of filtering on RefCOCO dataset using IoU as metric. n denotes number of filtering times. WordFilter means utilizing every word feature to generate kernel as DMN [28].

Model	n	val	testA	testB
LTS	1	65.43	67.76	63.08
	2	**66.04**	**68.68**	**63.27**
	3	65.54	67.82	62.97
LTS-WordFilter	1	64.92	67.31	62.50

Table 6. Results of LTS with different input resolutions on the RefCOCO dataset using IoU as metric.

Model	resolution	val	testA	testB
LTS (n=1)	320×320	63.01	65.40	60.78
	352×352	64.04	66.24	61.76
	384×384	64.45	67.02	62.47
	416×416	65.43	67.76	63.08
	448×448	65.71	67.90	63.23
	480×480	**65.90**	**68.16**	**63.45**

Table 7. Results of utilizing transformer, more filters and larger input resolution on the val sets of three datasets.

Model	RecCOCO	RecCOCO+	RecCOCOg
LTS	65.43	54.21	54.40
LTS*	**66.75**	**54.94**	**54.51**

feature with visual feature to build multi-modal representations. It can be seen that the IoU accuracy on val dataset drops 2.4% (Fusion+Filter vs Filter) and 1.4% (Fusion+Filter+Seg vs Filter+Seg). The fusion module proves the effectiveness of multi-modal representation in learning the semantic alignment between visual and linguistic modalities. Then we investigate the importance of relevance filtering by removing it from Fusion+Filter and Fusion+Filter+Seg. The IoU accuracy on val dataset drops 1.9% and 1.5%, respectively, which demonstrates that obtaining the location of the referent by the language description is beneficial to enhance the segmentation results. Comparing the result between Fusion (Filter / Fusion+Filter) and Fusion+Seg (Filter+Seg / Fusion+Filter+Seg), we can find that segmentation module can effectively improve the performances by obtaining a refined segmentation mask. In addition, we add this proposed segmentation module on CMPC [13] as shown in Tab. 3. The improved performances demonstrate that our module can generate more precise prediction mask. Finally, we can see that adding the locating loss also obtains better performance by supervising the alignments between image and text.

Tab. 4 shows the results when utilizing transformer instead of filtering as the localization module. Using more complex attention model can further improve the performance by locating the referent better.

Tab. 5 shows the experimental results when adopting multiple relevance filters, where $n = 2$ means we utilize relevance filtering twice in our model. When $n = 2$, our method gets better performance (+0.61 IoU). Such score is much better than previous published best result. For simplicity, all other experiments are performed with $n = 1$. Besides, we conduct an experiment by utilizing every word to generate kernel as DMN [28]. The results are shown in Tab. 5, where they obtain comparative performances. Considering the simplicity, we adopt sentence-based filtering in this paper.

In addition, we find that larger input resolution will improve the performance by providing richer information as shown in Tab. 6. In this paper, we set the image to 416×416 for fair comparison with previous methods.

Furthermore, we perform the experiments with the setting of using transformer, more filtering times, and larger input resolution on RefCOCO, RefCOCO+, and RefCOCOg datasets. The results are shown in Tab. 7. We can see that our model obtains better performances with this setting.

5. Conclusion

Referring image segmentation is a challenging task since it not only needs to discover the most relevant area given a language query, but also generate accurate object segmentation masks. In this work, we have developed a simple yet effective method for this task. Our approach decouples this task into two sub-sequential tasks: referring object prior prediction and fine object segmentation mask generation. Through explicitly modeling the position prior, we get much higher segmentation performance compared with previous best results. Extensive ablation studies verify the effectiveness of each component of our method.

Although the IoUs of our method are much higher than previous works, the mask quality is far from ground-truth (Fig. 5). We believe recent progress on image segmentation such as rendering [17] could give better mask quality. Besides, we only utilized simple visual and linguistic feature extraction backbones. More complex network structures have the potential to further improve the performance.

6. Acknowledgments

This work is supported by National Natural Science Foundation of China (61976214, 61721004, 61633021).

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 2

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Int. Conf. Comput. Vis.*, 2019. 4

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Eur. Conf. Comput. Vis.*, 2016. 3

[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.*, 2019. 2, 3

[5] 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.*, 2017. 2, 4

[6] Yi Wen Chen, Yi Hsuan Tsai, Tiantian Wang, Yen Yu Lin, and Ming Hsuan Yang. Referring expression object segmentation with caption-aware consistency. In *Brit. Mach. Vis. Conf.*, 2019. 5, 6

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. 3

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Int. Conf. Comput. Vis.*, 2017. 2

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 2

[10] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[11] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Eur. Conf. Comput. Vis.*, 2016. 1, 2

[12] 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.*, 2020. 6

[13] 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.*, 2020. 2, 5, 6, 7, 8

[14] 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.*, 2020. 2, 5, 6

[15] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 2, 5

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014. 5

[17] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 8

[18] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Adv. Neural Inform. Process. Syst.*, 2011. 5, 6

[19] 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.*, 2018. 2, 6

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[21] 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 *Eur. Conf. Comput. Vis.*, 2014. 5

[22] 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.*, 2017. 2, 6

[23] 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.*, 2019. 6

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[25] 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*, 2020. 5, 6

[26] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 5, 6, 7

[27] 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.*, 2016. 2, 5

[28] 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.*, 2018. 2, 3, 5, 6, 8

[29] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Eur. Conf. Comput. Vis.*, 2016. 5

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014. 3

[31] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[32] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Eur. Conf. Comput. Vis.*, 2018. 2, 3

[33] 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.*, 2017. 3

[34] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[35] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. SOLO: Segmenting objects by locations. In *Eur. Conf. Comput. Vis.*, 2020. 2

[36] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[37] 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.*, 2019. 1, 2, 3, 5, 6

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 6

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