Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation

Guang Feng[1,2], Zhiwei Hu[1,2], Lihe Zhang[1,2,†], Huchuan Lu[1,2]
[1]School of Information and Communication Engineering, Dalian University of Technology
[2]Ningbo Institute, Dalian University of Technology
fengguang.gg@gmail.com, hzw950822@mail.dlut.edu.cn,
{zhanglihe, lhchuan}@dlut.edu.cn

Abstract

Recently, referring image segmentation has aroused widespread interest. Previous methods perform the multi-modal fusion between language and vision at the decoding side of the network. And, linguistic feature interacts with visual feature of each scale separately, which ignores the continuous guidance of language to multi-scale visual features. In this work, we propose an encoder fusion network (EFN), which transforms the visual encoder into a multi-modal feature learning network, and uses language to refine the multi-modal features progressively. Moreover, a co-attention mechanism is embedded in the EFN to realize the parallel update of multi-modal features, which can promote the consistent of the cross-modal information representation in the semantic space. Finally, we propose a boundary enhancement module (BEM) to make the network pay more attention to the fine structure. The experiment results on four benchmark datasets demonstrate that the proposed approach achieves the state-of-the-art performance under different evaluation metrics without any post-processing.

1. Introduction

Referring image segmentation aims to extract the most relevant visual region (object or stuff) in an image based on the referring expression. Unlike the traditional semantic and instance segmentation, which require to correctly segment each semantic category or each object in an image, referring image segmentation needs to find a certain part of the image according to the understanding of the given language query. Therefore, it can be regarded as a pixel-wise foreground/background segmentation problem, and the output result is not limited by the predefined semantic categories or object classes. This task has a wide range of potential applications in language-based human-robot interaction.

†Corresponding Author

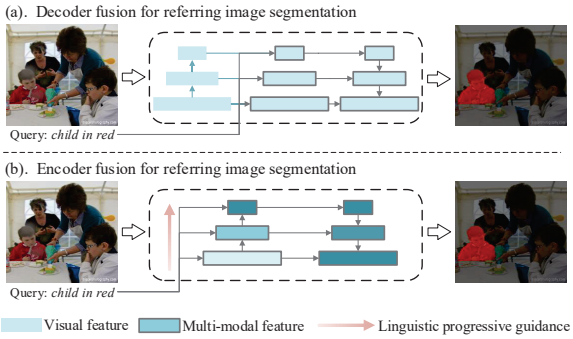


Figure 1: Two multi-modal fusion mechanisms. Existing methods achieve the fusion between language and vision in the decoder, while the proposed method does it in the encoder.

The key of this task is to realize the cross-modal matching between visual and linguistic features. The deep-learning community has rapidly improved the results of vision-language tasks over a short period of time. The rapid development of convolutional neural network (CNN) and recurrent neural network (RNN) have made a qualitative leap in the ability of understanding vision and language, thereby they can solve more complex pixel-level cross-modal prediction tasks. Early referring image segmentation methods [14, 26, 23, 33] mainly rely on the powerful learning ability of deep learning model. They directly concatenate linguistic features with visual features of each region, and then use the combined multi-modal features to generate the segmentation mask. Due to the lack of sufficient interaction between two modalities, such solutions can not meet the requirements of real-world applications. Recently, some works [36, 38, 1, 16, 17, 19] began to consider the linguistic and visual attention mechanisms to better aggregate these two kinds of features.

Although some referring image segmentation methods have been proposed in the last few years, there are still many problems that have not been explored. On the one hand, for the cross-modal fusion of vision and language. Previ-

ous methods usually adopt the **decoder fusion** strategy, in which the RGB image and the referring expression are fed into CNN or RNN to generate their own feature representations separately, and then fuse these features in the decoding stage. However, this fusion strategy at the output side of the network either only considers the interaction between linguistic and highest-level visual features [26, 23] or combines the linguistic features with the visual features of each level independently (as shown in Fig. 1 (a)) [38, 16, 19]. They do not investigate the deep guidance of language to multi-modal fused features. Besides, some works utilize visual and linguistic attention mechanisms for cross-modal feature matching. But they update the linguistic and visual features in a serial mode [36, 1, 16, 17, 19], that is, they only update the feature of one modality at a specific time, which will lead to the update delay of the features between different modalities and eventually weaken the consistency of the representation of multi-modal information. On the other hand, in CNNs, the repeated stride and pooling operations may lead to the loss of some important fine-structure information, but few referring image segmentation methods explicitly consider the problem of detail recovery.

To resolve the aforementioned problems, we propose an encoder fusion network with co-attention embedding (CEFNet) for referring image segmentation. Instead of the cross-modal information fusion at the output side, we adopt the **encoder fusion** strategy for the first time to progressively guide the multi-level cross-modal features by language. The original visual feature encoder (e.g., ResNet) is transformed into a multi-modal feature encoder (as shown in Fig. 1 (b)). The features of two modalities are deeply interleaved in the CNN encoder. Furthermore, to effectively play the guiding role of language, we adopt the co-attention mechanism to simultaneously update the features of different modalities. It utilizes the same affinity matrix to project different features to the common feature subspace in a parallel mode and better achieve the cross-modal matching to bridge the gap between coarse-grained referring expression and highly localized visual segmentation. We implement two simple and effective co-attention mechanisms such as vanilla co-attention and asymmetric co-attention, which offer a more insightful glimpse into the task of referring image segmentation. Finally, we design a boundary enhancement module (BEM), which captures and exploits boundary cues as guidance to gradually recover the details of the targeted region in the decoding stage of the network.

Our main contributions are as follows:

- We propose an encoder fusion network (EFN) that uses language to guide the multi-modal feature learning, thereby realizing deep interweaving between multi-modal features. In the EFN, the co-attention mechanism is embedded to guarantee the semantic alignment of different modalities, which promotes the representation ability of the language-targeted visual features.
- We introduce a boundary enhancement module (BEM) to emphasize the attention of the network to the contour representation, which can help the network to gradually recover the finer details.
- The proposed method achieves the state-of-the-art performance on four large-scale datasets including the UNC, UNC+, Google-Ref and ReferIt with the speed of 50 FPS on an Nvidia GTX 1080Ti GPU.

2. Related Work

Semantic and Instance Segmentation. The former aims at grouping pixels in a semantically meaningful way without differentiating each instance. The latter requires to separate all instances of objects rather than the stuff. In recent years, many semantic segmentation methods adopt fully convolutional network (FCN) [29] for end-to-end prediction. On this basis, the multi-scale context [43, 3, 4, 8, 10] and attention mechanisms [44, 11, 45, 18] are deeply examined. Some works [34, 8] leverage the encoder-decoder structures to alleviate the loss of details caused by continuous downsampling. Also, RGB-D based methods [12, 5] introduce depth prior to improve the performance. These methods provide inspirations for referring image segmentation.

In instance segmentation, Mask-RCNN [13] is a classical framework, which uses two-stage design to sequentially generates proposals and classify/segment them. In the follow-up works, feature pyramid [25], top-down and bottom-up [28], iterative optimization [2] and boundary-aware mechanisms [7] are explored. The success of boundary refinement strategy provides us with an important clue to solve the problem of referring image segmentation.

Referring Image Comprehension. This task has two branches: localization and segmentation. For referring image localization, previous methods are mainly composed of two separate stages. They firstly use object detector to extract candidate regions, and then rank these regions according to the referring expression. Pioneering methods [15, 32, 31] use the CNN-LSTM structure to select the object with the maximum posterior probability of the expression, and other works [27, 41] optimize the joint probability of the target object and the expression. Recently, some methods [37, 35, 24] use a one-stage framework. Instead of generating excessive candidate boxes, they directly predict the coordinates of the targeted region in an end-to-end manner. The above methods all implement the multi-modal fusion in the decoder.

For referring image segmentation, early methods [14, 26, 23, 33] directly concatenate language and visual features and then completely depend on a fully convolutional network to infer the pixel-wise mask. These methods

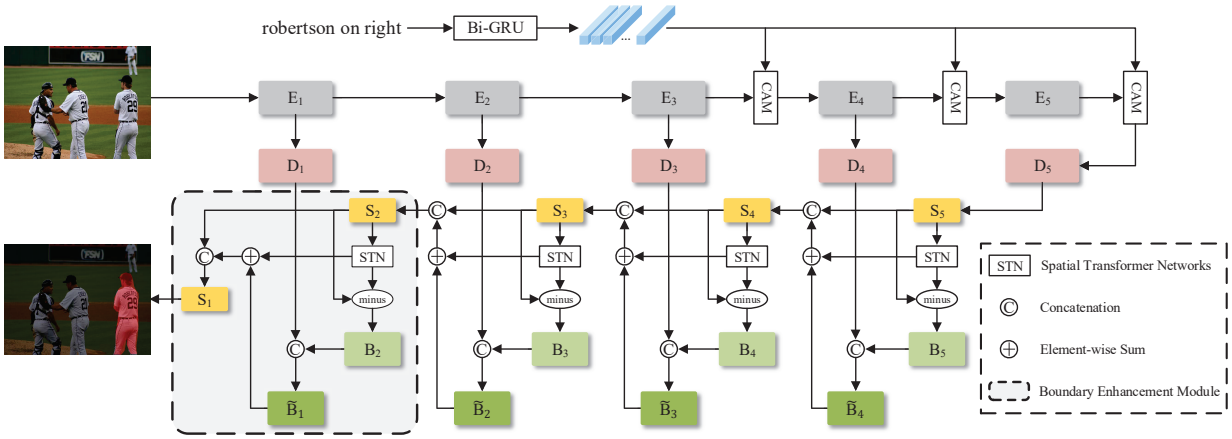


Figure 2: The overall architecture of our model. It mainly consists of the Bi-GRU encoder, ResNet-101 encoder ($E_1 \sim E_5$), co-attention module (CAM), decoder blocks ($D_1 \sim D_5$) and boundary enhancement module (BEM) assembled at the decoding end. The CAM is used to realize the matching between multi-modal features. The BEM captures the boundary cues and uses them to recover the details of the image, which produces a more accurate segmentation mask. The details of the proposed method are introduced in Sec. 3

do not explicitly formulate the intra-modal and inter-modal relationships. Some recent works [36, 38, 1, 16, 17, 19] consider the self-attention and cross-attention mechanisms of linguistic and visual information. For example, Shi et al. [36] adapt the vision-guided linguistic attention to learn the adaptive linguistic context of each visual region. Ye et al. [38] employ multiple non-local modules to update each pixel-word mixed features in a fully-connected manner. Hu et al. [16] design a bi-directional relationship inferring network to model the relationship between language and vision, which realizes the serial mutual guidance between multi-modal features. Huang et al. [17] firstly perceives all the entities in the image according to the entity and attribute words, and then use the relational words to model the relationships of all entities. LSCM [19] utilizes word graph based on dependency parsing tree to guide the learning of multi-modal context. Similarly, these methods also use the decoder fusion strategy. In addition, they do not update linguistic and visual features in parallel, which may weaken the consistency of language and vision in the semantic space. Different from the previous works, we design a parallel update mechanism to enhance the compatibility of multi-modal representation, and the multi-modal feature matching is performed in the encoder. We also propose a boundary enhancement module to guide the progressive fusion of multi-level features in the decoding stage.

3. Proposed Method

The overall architecture of the proposed method is illustrated in Fig. 2. In this section, we mainly introduce the co-attention based encoder fusion network and the boundary enhanced decoder network.

3.1. Encoder Fusion with Co-Attention

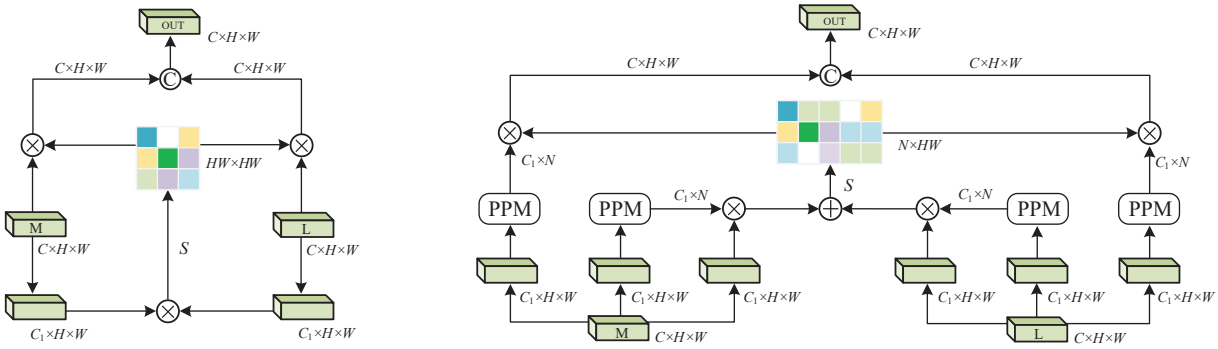
Encoder fusion network. For a input image, we use ResNet101 [42] to extract visual features. The ResNet101 is composed of five basic blocks: $conv1$, $res2$, $res3$, $res4$, and $res5$. The feature maps from these five blocks are represented as $\{\mathrm{E}_i\}_{i=1}^{5}$. To avoid losing excessive spatial details, the stride of the last block is set to 1. Unlike previous method that performs multi-modal fusion in the decoder, we insert language features after $res3$, $res4$, and $res5$ respectively. ResNet is converted into a multi-modal feature extractor. This design takes full advantage of the data fitting ability of **deep** CNN model and realizes the deep interleaving of cross-modal features. The experimental comparison between encoder fusion network (EFN) and decoder fusion network (DFN) is implemented, and the results on the UNC dataset are shown in Tab. 3.

Multi-modal feature representation. For a given expression, we feed the word embeddings $\{e_t\}_{t=1}^{T}$ into the Bi-GRU to generate the linguistic context $\{h_t\}_{t=1}^{T}$, where T represents the length of the language. Besides, we adopt a simple concatenation strategy to generate the initial multi-modal feature and denote it as:

$$m_p = w[e_i^p, h_T, s_i^p], \tag{1}$$

where e_i^p is a feature vector of E_i at position p. s_i^p represents 8-D spatial coordinates, which follows the design in [16]. w is the learnable parameters. Then we use m_p to calculate the position-specific linguistic context l_p:

$$\begin{aligned} \alpha_{p,t} &= {m_p}^{\top} \cdot e_t, \\ l_p &= \sum_{t=1}^{T} e_t \cdot \frac{\exp(\alpha_{p,t})}{\sum_{t=1}^{T} \exp(\alpha_{p,t})}. \end{aligned} \tag{2}$$



(a). Vanilla co-attention

(b). Asymmetric co-attention

Figure 3: Two co-attention modules. M: Initial multi-modal features. L: Adaptive linguistic context. S: Softmax. PPM: Pyramid pooling module. Ⓒ: Concatenation. ⊗: Matrix multiplication. ⊕: Element-wise summation. C, H and W are channel number, height and width of feature maps, respectively.

l_p treats each word differently. It can suppress the noise in the language expression and highlight the desired region. Next, the feature maps $\mathrm{M} = [m_p]$ and $\mathrm{L} = [l_p]$ go through the co-attention module to achieve the multi-modal fusion.

Vanilla co-attention. We design a co-attention scheme, which can model the dependencies between multi-modal features and project the multi-modal features to the common feature subspace. For the convenience of description, the size of M is defined as $C \times H \times W$, where H, W and C represent its height, width and channel number, respectively. The feature L has the same dimension as the M. At first, the features M and L are flattened into matrix representations with size $C \times (HW)$. Their affinity matrix $\mathrm{A} \in \mathbb{R}^{HW \times HW}$ is calculated as follows:

$$\mathrm{A} = (\mathrm{W}_m \mathrm{M})^\top (\mathrm{W}_l \mathrm{L}), \tag{3}$$

where $\mathrm{W}_m, \mathrm{W}_l \in \mathbb{R}^{C_1 \times C}$ are the learnable parameters. The element $a_{i,j}$ of A represents the similarity between the i^{th} position of M and the j^{th} position of L.

Then, we use the softmax function to normalize the similarity matrix as follows:

$$\begin{aligned} \mathrm{A}_1 &= \mathrm{softmax}(\mathrm{A}), \\ \mathrm{A}_2 &= \mathrm{softmax}(\mathrm{A}^\top), \end{aligned} \tag{4}$$

where A_1 and A_2 are the results of the row-wise and column-wise normalization, respectively. Thus, the feature maps M and L can be updated through weighted summing:

$$\begin{aligned} \widetilde{\mathrm{M}} &= \mathrm{M}\mathrm{A}_1{}^\top, \\ \widetilde{\mathrm{L}} &= \mathrm{L}\mathrm{A}_2{}^\top. \end{aligned} \tag{5}$$

We concatenate $\widetilde{\mathrm{M}}$ and $\widetilde{\mathrm{L}}$ along channel dimension and follow with a 3×3 convolution to get the multi-modal features $\mathrm{F} \in \mathbb{R}^{C_2 \times H \times W}$. The F is normalized and added to the encoder feature E. Thus, the embedding of the multi-modal feature in the encoder is finished. This mechanism can provide extra complementary cues according to the information of the other modality to implement the mutual guidance between these two modalities. Fig. 3 (a) shows the detailed structure of vanilla co-attention module (VCM).

Asymmetric co-attention. Furthermore, we propose an asymmetric co-attention module (ACM) to reduce the computational cost. Inspired by [45], we employ pyramid pooling module (PPM) to sample the feature maps M and L. The PPM is composed of four-scale feature bins, which are then flattened and concatenated to form a matrix of size $C_1 \times N$, $N \ll HW$. Here, the sizes of the feature bins are set to 1×1, 3×3, 6×6 and 8×8, respectively. Thus, the self-affinity matrixes of M and L can be calculated as:

$$\begin{aligned} \mathrm{SA}_m &= (\mathrm{PPM}(\mathrm{W}_m^1 \mathrm{M}))^\top (\mathrm{W}_m^2 \mathrm{M}), \\ \mathrm{SA}_l &= (\mathrm{PPM}(\mathrm{W}_l^1 \mathrm{L}))^\top (\mathrm{W}_l^2 \mathrm{L}), \end{aligned} \tag{6}$$

where SA_m and SA_l denote the modality-specific similarity matrixes. Their sizes are fixed to $N \times (HW)$ through the PPM, which is asymmetric. W_m^1, W_m^2, W_l^1 and W_l^2 indicate the learnable parameters. We further combine these two matrices as follows:

$$\mathrm{A}_3 = \mathrm{softmax}((\mathrm{SA}_m + \mathrm{SA}_l)^\top). \tag{7}$$

Then, the row-wise normalized matrix $\mathrm{A}_3 \in \mathbb{R}^{(HW) \times N}$ is used to assist the update of multi-modal features:

$$\begin{aligned} \widetilde{\mathrm{M}} &= \mathrm{A}_3(\mathrm{PPM}(\mathrm{W}_m^3 \mathrm{M}))^\top, \\ \widetilde{\mathrm{L}} &= \mathrm{A}_3(\mathrm{PPM}(\mathrm{W}_l^3 \mathrm{L}))^\top. \end{aligned} \tag{8}$$

Similarly to the vanilla co-attention, $\widetilde{\mathrm{M}}$ and $\widetilde{\mathrm{L}}$ is concatenated to generate the final multi-modal output. The whole structure of ACM is shown in Fig. 3 (b).

3.2. Boundary Enhancement Module

In CNNs, the repeated stride and pooling operations lead to the loss of fine structure information, which may blur the contour of the predicted region. Previous works [38, 1, 16, 17, 19] do not explicitly consider the restoration of details when performing multi-scale fusion in decoder. In this work, we design a boundary enhancement module (BEM), which uses boundary features as a guidance to make the network attend to finer details and realize the progressive refinement of the prediction. Its structure is shown in Fig. 2. Specifically, for the decoder features $\{\mathrm{D}_i\}_{i=1}^{5}$, we first compute the boundary-aware features:

$$\mathrm{B}_i = \mathrm{S}_i - \mathrm{STN}(\mathrm{S}_i), \tag{9}$$

where STN represents a spatial transformer networks [20]. Here, we utilize it to sample the high-level abstract semantic information from S_i. Thus, the residual B_i describes the fine structure. The prediction process of the boundary map can be written as:

$$\begin{aligned} \widetilde{\mathrm{B}}_{i-1} &= Conv(Cat(\mathrm{B}_i, \mathrm{D}_{i-1})), \\ \mathbf{BM}_{i-1} &= Sig(Conv(\widetilde{\mathrm{B}}_{i-1})), \end{aligned} \tag{10}$$

where $Cat(\cdot,\cdot)$ is the concatenation operation along the channel axis. $Conv$ and Sig denote the convolutional layer and sigmoid function, respectively. $\mathbf{BM}_{i-1}$ is supervised by the ground-truth contour of the targeted region.

Next, we exploit boundary feature $\widetilde{\mathrm{B}}_{i-1}$ to refine the segmentation mask as follows:

$$\begin{aligned} \mathrm{S}_{i-1} &= Conv(Cat(\widetilde{\mathrm{B}}_{i-1} + \mathrm{STN}(\mathrm{S}_i), \mathrm{S}_i)), \\ \mathbf{SM}_{i-1} &= Sig(Conv(\mathrm{S}_{i-1})), \end{aligned} \tag{11}$$

where S_{i-1} actually combines the information of decoder features D_i and D_{i-1}. $\mathbf{SM}_{i-1}$ denotes the refined mask, which is supervised by the ground-truth segmentation. The $\mathbf{SM}_1$ from the last decoder block is taken as the final prediction map, as illustrated in Fig. 2.

4. Experiments

4.1. Datasets

To verify the effectiveness of the proposed method, we evaluate the performance on four datasets, which are the UNC [40], UNC+ [40], Google-Ref [32] and ReferIt [21].

UNC: It contains 19,994 images with 142,209 language expressions for 50,000 segmented object regions. These data are selected from the MS COCO dataset using a two-player game [21]. There are multiple objects with the same category in each image.

UNC+: It is also a subset of the MS COCO, which contains 141,564 language expressions for 49,856 objects in 19,992 images. However, the referring expression does not contain the words that indicate location information, which means that the matching of their language and visual region totally depend on the appearance information.

Table 2: IoU for different length referring expressions on Google-Ref, UNC, UNC+ and ReferItGame.

	Length	1-5	6-7	8-10	11-20
G-Ref	R+LSTM [26]	32.29	28.27	27.33	26.61
	R+RMI [26]	35.34	31.76	30.66	30.56
	BRINet [16]	51.93	47.55	46.33	46.49
	Ours(VCM)	57.96	52.19	48.78	46.67
	Ours(ACM)	59.92	52.94	49.56	46.21

	Length	1-2	3	4-5	6-20
UNC	R+LSTM [26]	43.66	40.60	33.98	24.91
	R+RMI [26]	44.51	41.86	35.05	25.95
	BRINet [16]	65.99	64.83	56.97	45.65
	Ours(VCM)	68.18	66.14	56.82	46.01
	Ours(ACM)	68.73	65.58	57.32	45.90

	Length	1-2	3	4-5	6-20
UNC+	R+LSTM [26]	34.40	24.04	19.31	12.30
	R+RMI [26]	35.72	25.41	21.73	14.37
	BRINet [16]	59.12	46.89	40.57	31.32
	Ours(VCM)	60.87	48.88	43.79	29.45
	Ours(ACM)	61.62	52.18	43.46	31.52

	Length	1	2	3-4	5-20
ReferIt	R+LSTM [26]	67.64	52.26	44.87	33.81
	R+RMI [26]	68.11	52.73	45.69	34.53
	BRINet [16]	75.28	62.62	56.14	44.40
	Ours(VCM)	77.73	66.02	59.74	45.75
	Ours(ACM)	78.19	66.63	60.30	46.18

Google-Ref: It includes 104,560 referring expressions for 54,822 objects in 26,711 images. The annotations are based on Mechanical Turk instead of using a two-player game. The average length of referring expressions in this dataset is 8.43 words.

ReferIt: It is collected from the IAPR TC-12 [9]. It is composed of 130,525 referring expressions for 96,654 object regions in 19,894 natural images. In addition, their annotations contain objects or stuff, and the expressions are usually shorter and more succinct than the other datasets.

4.2. Implementation Details

The proposed framework is built on the public pytorch toolbox and is trained on an Nvidia GTX 1080Ti GPU for 200,000 iterations. Our network is trained by an end-to-end strategy and using the SGD optimizer with an initial learning rate of 0.00075 and divided by 10 after 100,000 iterations. All input images are resized to 320×320. The

Table 1: Quantitative evaluation of different methods on four datasets. -: no data available. DCRF: DenseCRF [22] post-processing.

*	ReferIt	UNC			UNC+			G-Ref
	test	val	testA	testB	val	testA	testB	val
LSTM-CNN$_{16}$ [14]	48.03	-	-	-	-	-	-	28.14
RMI+DCRF$_{17}$ [26]	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
DMN$_{18}$ [33]	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
KWA$_{18}$ [36]	59.19	-	-	-	-	-	-	36.92
RRN+DCRF$_{18}$ [23]	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
MAttNet$_{18}$ [39]	-	56.51	62.37	51.70	46.67	52.39	40.08	-
lang2seg$_{19}$ [6]	-	58.90	61.77	53.81	-	-	-	-
CMSA+DCRF$_{19}$ [38]	63.80	58.32	60.61	55.09	43.76	47.60	37.89	39.98
STEP$_{19}$ [1]	64.13	60.04	63.46	57.97	48.19	52.33	40.41	46.40
CGAN$_{20}$ [30]	-	59.25	62.37	53.94	46.16	51.37	38.24	46.54
BRINet+DCRF$_{20}$ [16]	63.46	61.35	63.37	59.57	48.57	52.87	42.13	48.04
LSCM+DCRF$_{20}$ [19]	66.57	61.47	64.99	59.55	49.34	53.12	**43.50**	48.05
CMPC+DCRF$_{20}$ [17]	65.53	61.36	64.54	59.64	49.56	53.44	43.23	49.05
Ours(VCM)	66.06	62.53	65.36	59.19	50.24	55.04	41.68	51.22
Ours(ACM)	**66.70**	**62.76**	**65.69**	**59.67**	**51.50**	**55.24**	43.01	**51.93**
Ours$^{\text{coco}}_{\text{VCM}}$	-	69.27	70.56	66.36	57.46	61.75	50.66	57.51
Ours$^{\text{coco}}_{\text{ACM}}$	-	68.97	71.13	66.95	57.48	61.35	51.97	57.49

Figure 4: Visual examples of referring image segmentation by our method.

weight decay and batch size are 0.0005 and 12, respectively. And when training G-ref, we use the UNC model as a pre-training model to avoid over-fitting. During the inference phase, the prediction map is resized to the same resolution as the original image. The binary cross entropy loss is used to supervise the boundary map and segmentation map, In addition, we also use the ground-truth segmentation (GT) to supervise the output of the STN.

Evaluation Metrics: Following previous works [16, 17, 19], we employ Overall Intersection-over-Union (Overall IoU) and Prec@X to evaluate the segmentation accuracy. The Overall IoU metric represents the ratio of the total intersection regions and the total union regions between the predicted mask and the ground truth for all the test samples. The Prec@X metric calculates the percentage of the IoU score of the prediction mask in the test set that exceeds the threshold X, where X $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.3. Performance Comparison

To verify the effectiveness of the proposed model, we compare it with thirteen methods, which are the LSTM-CNN [14], RMI [26], DMN [33], KWA [36], RRN [23], MAttNet [39], lang2seg [6], CMSA [38], STEP [1], C-GAN [30], BRINet [16], LSCM [19], and CMPC [17].

Performance Evaluation: Tab. 1 shows the performance (IoU) comparison of different methods on four datasets, in which Our(VCM) and Our(ACM) represent the results of using vanilla co-attention module and asymmet-

Table 3: Ablation study on the UNC val, testA and testB datasets.

	DFN	EFN	VCM	ACM	BEM	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	overall IoU
val	✓					57.30	50.40	42.00	28.42	9.00	52.86
		✓				64.16	58.45	51.16	37.53	13.39	55.87
		✓	✓			68.61	63.02	54.55	40.20	13.67	59.65
		✓		✓		69.22	64.11	56.33	41.67	15.32	60.09
		✓	✓		✓	74.07	68.84	61.76	48.74	20.06	62.53
		✓		✓	✓	73.95	69.58	62.59	49.61	20.63	62.76
testA	✓					61.66	54.30	44.55	31.09	9.32	55.98
		✓				67.03	61.59	53.92	40.13	13.31	58.07
		✓	✓			72.14	66.80	58.21	43.03	13.19	62.10
		✓		✓		72.95	67.90	59.98	45.04	14.46	62.46
		✓	✓		✓	77.53	73.18	66.02	52.11	18.88	65.36
		✓		✓	✓	77.66	73.73	66.70	52.75	19.66	65.69
testB	✓					52.31	45.48	37.51	26.85	10.40	49.66
		✓				58.61	52.62	45.67	34.56	15.03	52.48
		✓	✓			64.53	57.96	50.54	37.94	16.39	56.76
		✓		✓		65.02	58.33	50.34	38.74	16.31	57.09
		✓	✓		✓	69.74	63.75	56.90	45.20	22.51	59.19
		✓		✓	✓	69.66	65.14	58.31	46.18	22.43	59.67

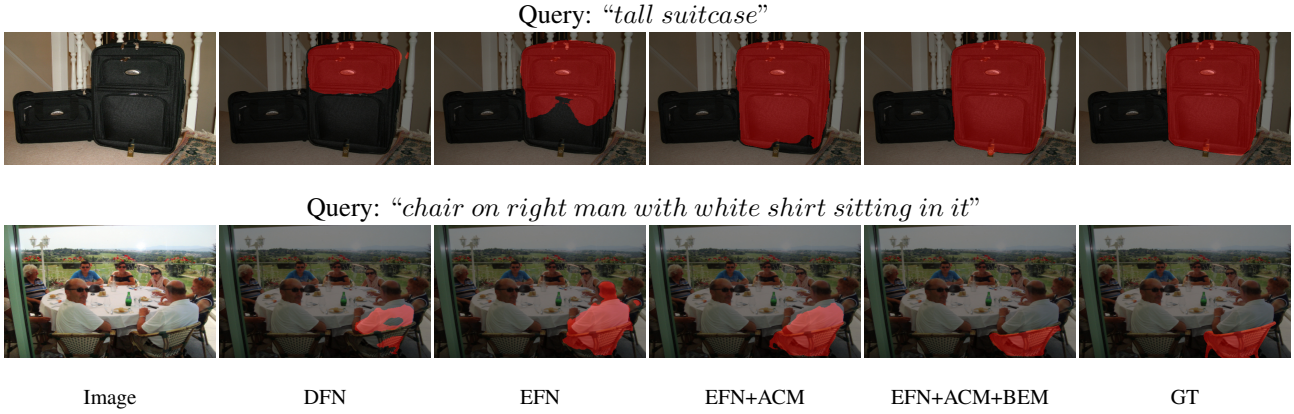


Figure 5: Visual examples of the proposed modules.

ric co-attention module, respectively. The proposed model consistently outperforms these competitors on most datasets except the UNC+ testB. Some methods like LSCM and CMPC apply DenseCRF [22] to refine their final masks while our model does not need any post-processing. In particular, we achieve the gain of 5.9%, 3.4% and 3.9% over the second best method CMPC [17] on the G-Ref, UNC+ testA and val, respectively. In addition, because the UNC, UNC+ and G-Ref are all collected from the MS COCO dataset, we combine their training data into a larger training set. The results of the model trained on it are denoted as $\text{Ours}^{\text{coco}}_{\text{VCM}}$ and $\text{Ours}^{\text{coco}}_{\text{ACM}}$, which show that sufficient training data can yield better results. We give some visual examples in Fig. 4. It can be seen that our method can accurately segment the specific regions (object or stuff) according to the query expression. Following [26, 16], we analyze the relationship between language length and segmentation accuracy. The results are demonstrated in Tab. 2, which indicate that our method achieves the state-of-the-art performance.

Runtime and Memory Statistics: We implement all the tests on a NVIDIA GTX 1080 Ti GPU. The comparison of running time is reported in Tab. 4. Our method runs the fastest with a speed of 50 FPS. The GPU memory usage is shown in Tab. 5. From Tab. 4 and Tab. 5, we can find that although the VCM has advantages in speed, the large input size causes the memory usage to sharply increase. On the contrary, the VCM is not sensitive to the input size. Therefore, it is widely applicable.

4.4. Ablation Study

we conduct a series of experiments on the UNC dataset to verify the benefit of each component.

Comparison of DFN and EFN: We first remove the co-attention module and boundary enhancement module from

Table 4: Runtime analysis of different methods. The time of post-processing is ignored.

	LSTM	RMI	RRN	CMSA	BRINet	CMPC	Ours(VCM)	Ours(ACM)
Time(ms)	58ms	72ms	43ms	79ms	117ms	60ms	17ms	20ms

Table 5: GPU memory (MB) comparisons between VCM and ACM. The lower values are the better.

Input size	512×20×20	512×40×40	512×96×96
VCM	9.93	54.35	1308.00
ACM	6.92	14.29	61.11

the CEFNet in Fig. 2. Then, we achieve the multi-modal fusion in the encoder by Eq. (1), and the decoder adopts the FPN [25] structure. This network is taken as the baseline network (EFN) of encoder fusion. In addition, similar to previous works, we realize the multi-modal fusion in the FPN decoder by Eq. (1) to build the baseline (DFN) for decoder fusion. We evaluate the two baselines in Tab. 3, from which we can see that EFN is significantly better than DFN. With the help of ResNet, the encoder fusion strategy achieves more powerful feature coding without increasing additional computational burden.

Effectiveness of Co-Attention: We evaluate the performance of the vanilla co-attention module (VCM) and asymmetric co-attention module(ACM). Compared with the baseline EFN, the VCM brings 6.8%, 6.9% and 8.2% IoU improvement on the UNC-val, UNC-testA, and UNC-testB, respectively. Similarly, the ACM achieves the gain of 7.6%, 7.6% and 8.8% on the same datasets, respectively. The ACM performs slightly better than the VCM. We attribute it to the modality-specific affinity learning, which focuses on important regions within the modality and achieves better contextual understanding of the modality itself. It is conducive to cross-modal alignment in the next stage.

Effectiveness of BEM: Tab. 3 presents the ablation results of boundary enhancement module (BEM), which show that the special consideration of boundary refinement can significantly improve the performance. BEM can bring about 2%∼3% performance improvement (Overall IoU)) to the final prediction result. Some visual results in Fig. 5 demonstrate the benefits of BEM. These figures show that the prediction mask can fit the object boundary more closely after the refinement of BEM.

4.5. Failure Cases

We visualize some interesting failure cases in Fig. 6. One type of failure occurs when queries are ambiguous. For example, for the top left example, the misspelling of a word ($roght \rightarrow right$) causes part of the semantics of the sentence to be lost. Also, for the top right example, there are two horse butts on the left. Another case is that when the query contains low-frequency or new words (e.g. in the bottom left example, *cop* rarely appears in the training data), our method sometimes fails to segment out the core region accurately. This problem may be alleviated by using one/zero-shot learning. Finally, we observed that sometimes small objects cannot be segmented completely (the bottom right example). This phenomenon can be alleviated by enlarging the scale of input images. Fortunately, the ACM is insensitive to the size (Tab. 5 for details).

Figure 6: Visual examples of the failure cases.

Through the analysis of successful (Fig. 4) and failure cases, we think that the co-attention module can learn the high-order cross-modal relationships even in some complicated semantic scenarios. It enables the network to pay more attention to the correlated, informative regions, and produce discriminative foreground features.

5. Conclusion

In this paper, we propose an encoder fusion network with co-attention embedding (CEFNet) to fuse multi-modal information for referring image segmentation. Compared with the decoder fusion strategy, our strategy adequately utilizes language to guide multi-model feature learning without increasing computational complexity. The designed co-attention module can promote the matching between multi-modal features and strengthen their targeting ability. Moreover, a boundary enhancement module is equipped to make the network pay more attention to the details. Extensive evaluations on four datasets demonstrate that the proposed approach outperforms previous state-of-the-art methods both in performance and speed. In future, we can extend our co-attention module to the one-stage grounding to promote the integration of language and vision.

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, the Dalian Science and Technology Innovation Foundation #2019J12GX039, and the Fundamental Research Funds for the Central Universities #DUT20ZD212.

References

[1] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Int. Conf. Comput. Vis.*, pages 7454–7463, 2019. 1, 2, 3, 5, 6

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4974–4983, 2019. 2

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

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[5] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. *arXiv preprint arXiv:2007.09183*, 2020. 2

[6] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang. Referring expression object segmentation with caption-aware consistency. In *Brit. Mach. Vis. Conf.*, 2019. 6

[7] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. *arXiv preprint arXiv:2007.08921*, 2020. 2

[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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2393–2402, 2018. 2

[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. *CVIU*, 114(4):419–428, 2010. 5

[10] Guang Feng, Hongguang Bo, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Cacnet: Salient object detection via context aggregation and contrast embedding. *Neurocomputing*, 403:33–44, 2020. 2

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3146–3154, 2019. 2

[12] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*, pages 213–228. Springer, 2016. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Eur. Conf. Comput. Vis.*, pages 108–124. Springer, 2016. 1, 2, 6

[15] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4555–4564, 2016. 2

[16] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4424–4433, 2020. 1, 2, 3, 5, 6, 7

[17] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10488–10497, 2020. 1, 2, 3, 5, 6, 7

[18] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 603–612, 2019. 2

[19] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. *arXiv preprint arXiv:2010.00515*, 2020. 1, 2, 3, 5, 6

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Adv. Neural Inform. Process. Syst.*, pages 2017–2025, 2015. 5

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 5

[22] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Adv. Neural Inform. Process. Syst.*, pages 109–117, 2011. 6, 7

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

[24] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10880–10889, 2020. 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017. 2, 8

[26] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Int. Conf. Comput. Vis.*, pages 1271–1280, 2017. 1, 2, 5, 6, 7

[27] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Int. Conf. Comput. Vis.*, pages 4856–4864, 2017. 2

[28] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE*

Conf. Comput. Vis. Pattern Recog., pages 8759–8768, 2018. 2

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

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

[31] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7102–7111, 2017. 2

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 2, 5

[33] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Eur. Conf. Comput. Vis.*, pages 630–645, 2018. 1, 2, 6

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

[35] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Int. Conf. Comput. Vis.*, pages 4694–4703, 2019. 2

[36] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Eur. Conf. Comput. Vis.*, pages 38–54, 2018. 1, 2, 3, 6

[37] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Int. Conf. Comput. Vis.*, pages 4683–4693, 2019. 2

[38] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10502–10511, 2019. 1, 2, 3, 5, 6

[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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1307–1315, 2018. 6

[40] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85. Springer, 2016. 5

[41] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7282–7290, 2017. 2

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3

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

[44] 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 *Eur. Conf. Comput. Vis.*, pages 267–283, 2018. 2

[45] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 593–602, 2019. 2, 4