This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Bidirectional Relationship Inferring Network for Referring Image Localization and Segmentation

Guang Feng, Zhiwei Hu, Lihe Zhang, *Member, IEEE*, Jiayu Sun, and Huchuan Lu, *Senior Member, IEEE*

***Abstract*— Recently, referring image localization and segmentation has aroused widespread interest. However, the existing methods lack a clear description of the interdependence between language and vision. To this end, we present a bidirectional relationship inferring network (BRINet) to effectively address the challenging tasks. Specifically, we first employ a vision-guided linguistic attention module to perceive the keywords corresponding to each image region. Then, language-guided visual attention adopts the learned adaptive language to guide the update of the visual features. Together, they form a bidirectional cross-modal attention module (BCAM) to achieve the mutual guidance between language and vision. They can help the network align the cross-modal features better. Based on the vanilla language-guided visual attention, we further design an asymmetric language-guided visual attention, which significantly reduces the computational cost by modeling the relationship between each pixel and each pooled subregion. In addition, a segmentation-guided bottom-up augmentation module (SBAM) is utilized to selectively combine multilevel information flow for object localization. Experiments show that our method outperforms other state-of-the-art methods on three referring image localization datasets and four referring image segmentation datasets.**

***Index Terms*— Language-guided visual attention, referring image localization and segmentation, segmentation-guided feature augmentation, vision-guided linguistic attention (VLAM).**

I. Introduction

BOTH referring image localization and referring image segmentation aim to mine the most relevant visual region in an image based on the referring expression. Their prediction results are box level and pixel level. Unlike traditional object detection or semantic segmentation, in which each instance or pixel needs to be assigned a specific semantic category label, referring image localization and segmentation tasks require a deeper understanding of the image. They need to consider appearance attributes, actions, spatial relationships, as well as semantic cues contained in the language expression. For example, if the expression is "a man sitting on the right is wearing a black suit," we need an algorithm that not only distinguishes all the instances in the image but also locates the most suitable one, according to the meaning of the sentence. The two tasks have a wide range of potential applications in language-based human–robot interaction.

With the development of deep learning technology, many complex cross-modal tasks are also possible to solve. Over the past few years, referring image localization and segmentation are usually explored and designed separately. For referring image localization, the existing methods either first extract all the object candidates and then rank them according to the referring expression (i.e., two-stage scheme) or directly predict a bounding box according to the input image and expression (i.e., one-stage scheme). The two-stage localization methods [2]–[10] require an explicit object detector to obtain object bounding boxes and their corresponding region of interest (ROI) features. However, the ground-truth regions of referring image localization may be any visual region, and the pretrained object detector fundamentally limits the possibility of localizing the target beyond the predefined domain. To avoid this problem, the one-stage methods [11]–[13] came into being. They divide the image into two parts (i.e., the foreground and background regions), and the region division depends on the guidance of language expression rather than the semantic category in the traditional sense. Their main ideas are to embed the features of language and vision into the one-stage object detector [14], [15] for end-to-end learning.

Benefiting from the powerful ability of deep neural networks, early referring image segmentation methods [16]–[19] directly use the "concatenate" operations to fuse the linguistic and visual features and then employ the fused results to generate the segmentation map directly. These methods indeed achieve good performance when dealing with simple referring expression or semantic scene. However, they do not consider the local context of image and language and lack the pixel-level interaction between cross-modal features. Later, CMSA [20] models the fully connected graph between each of vision word mixed features, and KWA [21] and STEP [22] employ the attention mechanism to reweight each word. However, they do not explicitly characterize the mutual guidance between the visual and linguistic features yet, thereby weakening the contextual consistency of linguistic and visual region in the feature space.

Manuscript received May 18, 2021; revised July 26, 2021; accepted August 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61876202 and Grant 61829102, in part by Dalian Science and Technology Innovation Foundation under Grant 2019J12GX039, and in part by the Fundamental Research Funds for the Central Universities under Grant DUT20ZD212. This article was presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020 [10.1109/CVPR42600.2020.00448]. *(Corresponding author: Lihe Zhang.)*

The authors are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: fengguang.gg@gmail.com; hzw950822@mail.dlut.edu.cn; zhanglihe@dlut.edu.cn; jiayusun666@gmail.com; lhchuan@dlut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3106153.

Digital Object Identifier 10.1109/TNNLS.2021.3106153

2162-237X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

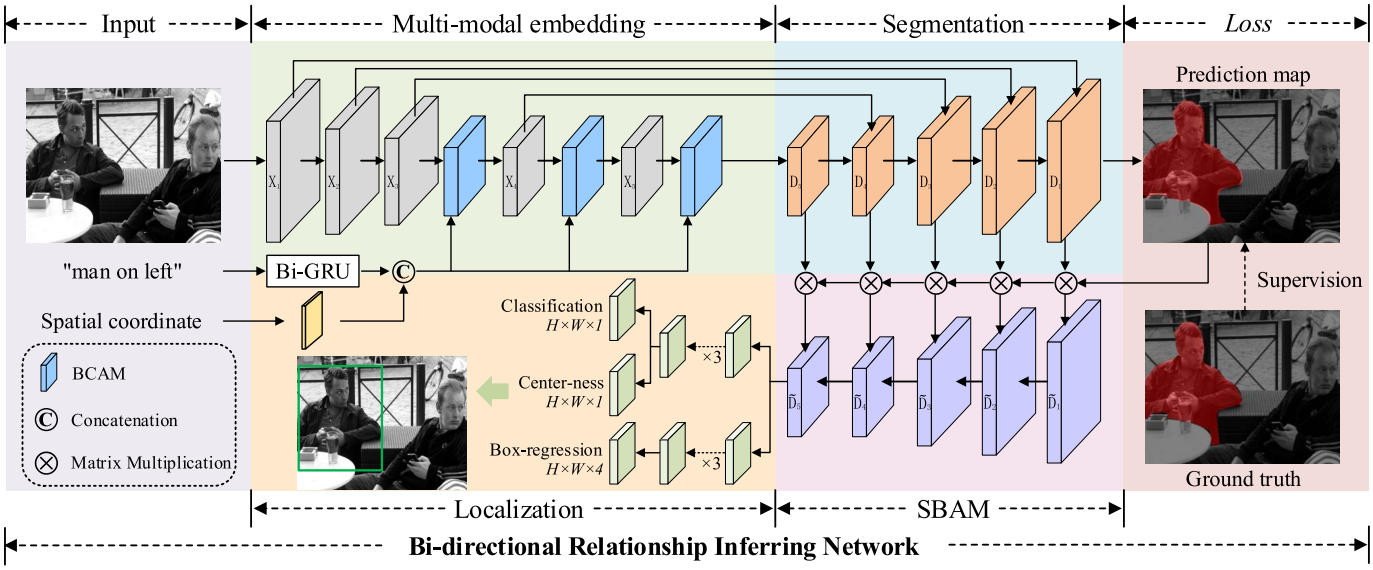

Fig. 1. Overall framework of our method. The image and text are fed into a multimodal embedding module to obtain the mixed features, in which the BCAM is used for cross-modal fusion. Then, the segmentation module and the localization module are used to produce the prediction mask and the bounding box, respectively. The segmentation-guided bottom-up feature augmentation module, denoted as SBAM, combines the multilevel features for localization.

The joint modeling of the two tasks is also examined in several works. The method [23] feeds the predicted bounding box to the mask branch of Mask R-CNN [24] to obtain a pixelwise mask, thereby achieving localization and segmentation of the relevant region, but its efficiency is severely limited by the two-stage localization. Recently, MCN [25] achieves the joint learning of referring image localization and segmentation based on a one-stage detection framework YOLOv3 [15], which mainly considers the cooperation between the two tasks and does not deeply investigate the interaction between multimodal features. To realize the mutual embedding of language and vision, thereby improving the performance of localization and segmentation simultaneously, we design a one-stage bidirectional relationship inferring network (BRINet) to effectively capture the dependencies of multimodal features under the guidance of both language and vision, and it predicts the pixel-level mask and bounding box in a serial manner.

First, we construct a vision-guided linguistic attention module (VLAM) to learn the adaptive linguistic context for each visual region. Second, a language-guided visual attention module (LVAM) utilizes the learned linguistic context to guide the learning of spatial dependencies between any two regions of the visual features. We implement two simple and effective LVAM such as vanilla LVAM (V-LVAM) and asymmetric LVAM (A-LVAM), which offer a more insightful glimpse for the task. VLAM and LVAM constitute the bidirectional cross-modal attention module (BCAM). By the mutual learning between different modalities, the proposed model enriches the contextual representation of the target region. Therefore, the target region can be highlighted more consistently with the help of referring expression. This apparently allows us to consider more complex and nonsequential dependencies between visual regions and words. In addition, we insert the BCAM into the encoder of the network (as shown in Fig. 1), which converts the CNN-based visual encoder into a multimodal feature encoder. The entire network no longer needs additional intermediate layers for multimodal fusion. Finally, we design a segmentation-guided bottom-up augmentation module (SBAM), which can selectively and progressively carry out multilevel feature aggregation to improve object localization.

A preliminary version of this work has appeared in [1]. In contrast to the conference work, we first extend the previous model into a multitask network, which can achieve efficient referring image localization and segmentation simultaneously. Second, we propose an A-LVAM, which can reduce the computational cost of modeling pixelwise relationship. It adopts spatial pyramid pooling to sample some subregions of different scales and then models the affinity between them and each pixel. This pixel-to-region operation reduces the size of the affinity matrix. Finally, we build a SBAM to select and aggregate the features with high response to the foreground at different scales, thereby enhancing the feature representation of target region for object localization.

Our main contributions are listed as follows.

1) We propose a new multitask network to achieve referring image segmentation and localization in a serial manner, which naturally employs the segmentation result to guide the localization. This is helpful for detecting stuff region.
2) We design an efficient BCAM, which utilizes the mutual guidance between language and vision to promote cross-modal fusion. It is equipped with the encoder for the early fusion of multimodal features to ensure the deep interaction between language and vision.
3) We introduce a SBAM to flexibly implement bottom-up cross-level feature fusion, which can further refine the bounding box of the referred entity.
4) The proposed BRINet achieves the state-of-the-art performance in both referring image localization and

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

referring image segmentation on four benchmark datasets. Also, our method runs faster than 55 FPS on a 1080Ti GPU.

II. Related Work

In this section, we summarize the recent research related to our work in the fields of semantic segmentation, instance segmentation, referring image localization, referring image segmentation, and attention mechanism.

A. Fully Convolutional Network for Localization and Segmentation

As fully convolutional networks (FCNs) [26] exhibit their powerful ability in handling pixelwise prediction in an end-to-end manner, many pixelwise prediction tasks have achieved tremendous successes in recent years. For the task of semantic segmentation, some of the state-of-the-art FCN-based methods are proposed to reduce downsampling degradation effects and reinforce context aggregation of multiple feature scales. Pyramid pooling module (PPM) is used in PSPNet [27] to perform a region-based collection of contexts in different scales. In [28] and [29], atrous spatial pyramid pooling (ASPP) has been proposed to fuse the multiscale contexts by expanding the receptive field. Meanwhile, some methods [30]–[33] investigate the autoencoder structure and complement the detail-rich low-level features back to the detail-missing high-level features for prediction. CCL [32] proposes a gated multiscale aggregation mechanism to control the information flow from high-level to low-level features. MCI [33] proposes to bidirectionally and recurrently combines the multiscale contexts by some interconnected LSTM chains. Besides, some RGB-D-based methods [34], [35] introduce depth map to supplement semantic information. Our method considers the fine-grained interaction between the linguistic and visual context for the cross-modal fusion.

For instance segmentation, Mask-RCNN [24] is a representative two-stage segmentation framework. It first generates candidate region-of-interests (ROIs) [36] and then classifies and segments these candidates. In the following works, feature pyramid [37], dual-path aggregation [38], and iterative optimization [39] are used to enhance the performance of the network. In addition, in order to pursue efficiency and avoid heavy head and multiple RoIs, some one-stage methods [40]–[44] that no longer need anchor as *a priori* are designed. Our network is a language-oriented one-stage referring image localization and segmentation method.

Recently, some tasks that combine vision and language have been proposed. For example, the goal of referring image localization is to localize the targeted object by given its corresponding natural language expression. The works [2], [3], [45] independently encode the inputs of the two modalities and then directly map the two features to the common feature space for matching. Some recent approaches [6], [7], [46] disassemble the expression into different elements and these elements are performed relationship modeling of each object. These aforementioned methods all rely on a predefined object detector to detect all candidate regions in the first stage and then rank them according to the referring expression. This process is very time-consuming. To solve this problem, some anchor-free methods [11]–[13] are proposed. Instead of generating candidate boxes, they design an end-to-end model to directly predict the referred region according to the input image and expression. FAOS [11] and ZSGNet [12] integrate linguistic features into a one-stage object detection framework (i.e., YOLOv3 [15] and SSD [47]), but they only use "concatenate" operation to realize the cross-modal fusion. RCCF [13] employs linguistic features as dynamic convolution kernel to realize the interaction between multimodal information. Obviously, these one-stage methods do not fully and deeply consider the multimodal interaction of language and vision.

Instead of generating a language-targeted bounding box of image region, the task of referring image segmentation predicts a precise segmentation mask. Hu *et al.* [16] first introduced this task. This method simply fuses both visual and linguistic features by concatenation for prediction. RMI [17] performs the region-based tiles of multimodal features and separately and sequential infer them for prediction by a two-layered LSTM network. To perform gradual fusion of the pyramid features, convolutional LSTM has been applied in RRN [18]. The word-specific features in multiple modalities are recurrently concatenated in [19]. These methods mentioned above characterize the linguistic and visual information by a "concatenation-convolution" procedure. However, the cross-modal relationship is still lack of explicit modeling. Later, the target object is highlighted in [21] by extracting key words and reducing the noise in the referring expression. In [20], self-attention mechanism is used to model the visual relationship of words in expression. Besides, image-to-word attention is utilized in [22], where the relationship heatmap between all words and visual regions is employed to recurrently guide the prediction. However, these methods only model the one-way relationship of different modalities.

Recently, MCN [25] presents a multitask collaborative network to learn referring image localization and segmentation simultaneously. However, this network mainly considers the mutual promotion between two tasks and does not carefully explore the deep embedding between language and vision. In our method, a bidirectional guidance mechanism is proposed to better adapt linguistic and visual features mutually.

B. Attention Mechanism

There has been a wide application of attention mechanism used in many kinds of tasks [5], [46], [48]–[59]. A co-attention mechanism is proposed in [5] to model the linguistic and visual feature relationship. Shi *et al.* [21] reweighted the relevance between each word and visual region by word attention. While Wang *et al.* [46] modeled interobject relationships by graph attention. In [7], a language-guided visual graph is proposed to model the relation among all the objects. In [60], question-adaptive attention is used to model the object relations in multiple types and learn an adaptive region representation. Unlike the previous works, in this work,

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

the proposed cross-modal attention mechanism is extended to both referring image localization and segmentation. Besides, bidirectional attention is designed to enhance the semantic consistency across different modalities.

III. Proposed Approach

The overall architecture of the proposed method is shown in Fig. 1. Given an image and its referring expression, we first use Darknet-53 [15] and Bi-GRU to extract visual and linguistic features, respectively. Then, the visual, linguistic, and spatial features are fed into the BCAM to model the relationships between multimodal features. These relationships are used to update the contextual representation of the target region. It is worth noting that BCAM is introduced into the encoder of the network. Thus, Darknet is transformed into a multimodal feature coding network and realizes the progressive early fusion of language and vision. The advantages of the encoder fusion strategy (EFS) over the traditional decoder fusion are verified in the experiments. Finally, the SBAM is designed to determine the ON-and-OFF of multilevel features and selectively pass useful information from the bottom layer to the top layer.

A. Multimodal Feature Encoding

1) Visual Feature Encoding: We use Darknet-53 [15] to construct the visual encoder, which contains five basic residual blocks. The features from these five blocks can be represented as $\{X_i\}_{i=1}^{5}$. Besides, in order to avoid losing excessive spatial details, the stride of the last block is set to 1.

2) Text Feature Encoding: For a given referring expression $R = \{r_t\}_{t=1}^{T}$, where t indexes the tth word and T is the length of expression. We first encode each word r_t into a real vector $e_t \in \mathbb{R}^{C_e}$ through the Bert embedding [61], and then, the Bi-GRU is used to encode the context of each word. Thus, the hidden state of r_t is the concatenation of its forward and backward hidden vectors, which is denoted as $\tilde{h}_t$. We employ a full connection layers to reduce the dimension of $\tilde{h}_t$, that is, $\tilde{h}_t \rightarrow h_t \in \mathbb{R}^{C_h}$.

3) Spatial Feature Encoding: Previous referring image localization and segmentation methods [1], [11], [20] usually concatenate spatial coordinate feature to enhance the spatial information. Following [11], we generate an 8-D spatial representation s_p at each position p, which describes the coordinates of the top-left corner, center, and bottom-right corner of each visual position.

B. Vision-Guided Linguistic Attention

For a given referring expression, there is a fact that the importance of each word in the sentence to each visual region is different. If we treat these language features equally and use them to guide cross-modal fusion directly, some noise may be introduced to make the network produce an erroneous prediction. Thus, we introduce a VLAM to adaptively establish the relationship between the linguistic context and each visual region. The relationship between the pth visual position and the tth word is defined as follows:

$$\begin{aligned} v_p &= w\left[x_p^i, h_T, s_p^i\right] \\ \alpha_{p,t} &= {v_p}^{\top} e_t \\ \widetilde{\alpha}_{p,t} &= \frac{\exp(\alpha_{p,t})}{\sum_{t=1}^{T}\exp(\alpha_{p,t})} \end{aligned} \tag{1}$$

where $x_p^i \in \mathbb{R}^{C_x}$ represents the visual feature vector of position p in X_i, $h_T \in \mathbb{R}^{C_h}$ denotes the final hidden state, and $s_p^i \in \mathbb{R}^8$ is the spatial coordinate feature. $[\cdot,\cdot]$ represents the concatenation operation. $w \in \mathbb{R}^{C_e \times (C_x + C_h + 8)}$ is the learnable parameter, which aims to map the mixed feature into the same dimension of the word embedding e_t. Before calculating attention, the tiled language feature is the representation of the whole sentence. In this way, the visual feature of every pixel is endowed with the same global context in advance. The pretiled sentence feature provides the global guidance for subsequent cross-modal inferring. $\widetilde{\alpha}_{p,t}$ is the normalized attention score, which represents the importance of the tth word to the pth feature region. Finally, the new linguistic context l_p for the pth feature region is calculated as follows:

$$l_p = \sum_{t=1}^{T} \widetilde{\alpha}_{p,t} e_t. \tag{2}$$

C. Vanilla Language-Guided Visual Attention

Contextual information is essential for referring image localization and segmentation, which helps the network locate and segment target region accurately. To model contextual relationship across different regions, we design a V-LVAM that leverages the region-adaptive linguistic features to compute the affinity between any two pixels.

For feature vector v_p, the normalized relationship weighted between itself and the qth region v_q is formulated as follows:

$$\begin{aligned} \lambda_{p,q} &= w_q[\tanh(w_l l_p + w_v v_p)] \\ \widetilde{\lambda}_{p,q} &= \frac{\exp(\lambda_{p,q})}{\sum_{q=1}^{N}\exp(\lambda_{p,q})} \end{aligned} \tag{3}$$

where $w_l \in \mathbb{R}^{C_1 \times C_e}$, $w_v \in \mathbb{R}^{C_1 \times C_e}$, and $w_q \in \mathbb{R}^{C_1}$ are the learnable parameters. N is the number of pixels. $\widetilde{\lambda}_{p,q}$ depicts the importance of the qth feature toward the pth feature region. Based on this processing, we establish the dependencies of all regions in the image. Then, we utilize these relevances to update the visual feature and then obtain the cross-modal feature f_p

$$\begin{aligned} \widetilde{v}_p &= \sum_{q=1}^{N}[\widetilde{\lambda}_{p,q}(w_{\widetilde{v}_p} v_q)] \\ f_p &= w_1[\widetilde{v}_p,\ l_p] + w_2 \widetilde{v}_p \end{aligned} \tag{4}$$

where $w_{\widetilde{v}_p}$, w_1, and w_2 are the parameters. Fig. 2(a) shows the detailed structure of V-LVAM, which is integrated with VLAM to constitute the BCAM.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

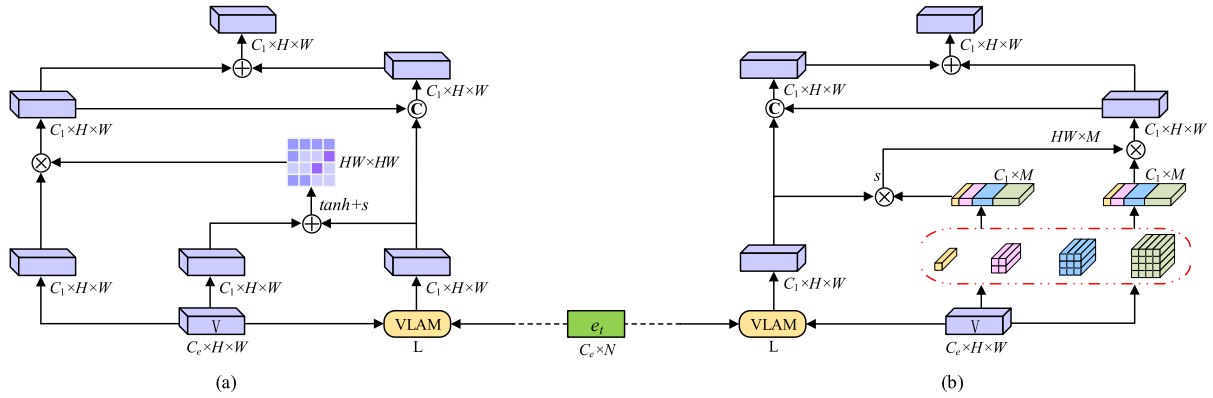

Fig. 2. BCAM. e_t: word embedding. s: softmax. VLAM: vision-guided linguistic attention module. L: adaptive linguistic context. V: visual feature. (a) Vanilla language-guided visual attention. (b) Asymmetric language-guided visual attention.

D. Asymmetric Language-Guided Visual Attention

The V-LVAM needs to calculate the relationship between any two pixels. To reduce the GPU utilization and computational cost of pixel-level operation, we further propose an A-LVAM to characterize the relationships between each pixel and a few numbers of pooled subregions. For the convenience of description, feature map $[v_p]_{p=1}^{N}$ and context map $[l_p]_{p=1}^{N}$ are denoted as V and L in the form of matrix, respectively. We use the PPM [27] to process them separately. PPM can fix them to a specific size through the pooling operation. In this work, we generate four pooled subregions with varied bin sizes of 1×1, 3×3, 6×6, and 8×8. Thus, the affinity between each pixel in L and each pooled subregion in V can be calculated as

$$A = (\mathrm{PPM}(W_1 V))^{\top}(W_2 L) \tag{5}$$

where $W_1, W_2 \in \mathbb{R}^{C_1\times C_e}$ are the learnable parameters. Equation (5) omits the flattening operation. The size of A is fixed to $M\times(HW)$ through the PPM, where M represents the number of subregions and $M \ll HW$. We further normalize affinity matrix A as follows:

$$\widetilde{A} = \mathrm{softmax}(A) \tag{6}$$

which is used to update visual feature map V

$$\widetilde{V} = \widetilde{A}^{\top}(\mathrm{PPM}(W_3 V)) \tag{7}$$

where W_3 is the parameter. In the process of update, we significantly reduce the size of affinity matrix from $HW\times HW$ to $M\times HW$. Similar to V-LVAM, $\widetilde{V}$ and L are fused, and then, the cross-modal feature is output

$$F = W_4[\widetilde{V},\ L] + W_5\widetilde{V} \tag{8}$$

where W_4 and W_5 are the parameters. The structure of A-LVAM is shown in Fig. 2(b).

E. Segmentation-Guided Bottom-Up Augmentation

To enhance the performance of object localization, we build a segmentation-guided feature augmentation module, which takes the segmentation mask as guidance and makes the network pay more attention to the target region of the image. Its detailed structure is shown in Fig. 1. We use the FPN [37] to compute the decoder features $\{D_i\}_{i=1}^{5}$ for referring image segmentation. The FPN realizes feature fusion from high level to low level. Then, we exploit SBAM to progressively combine features from shallow layer to deep layer for referring image localization. The newly generated feature maps are denoted as $\{\widetilde{D}_i\}_{i=1}^{5}$. In the bottom-up process, we use the segmentation map to control the message passing in the network and strengthen the representation of the context of the foreground. This process is formulated as

$$\widetilde{D}_i = W(\widetilde{D}_{i-1} + S\odot D_i) \tag{9}$$

where S represents the segmentation map, $\odot$ denotes the elementwise multiplication, and W is the learnable parameter.

F. Bounding Box and Mask Prediction

Referring image segmentation can be regarded as a foreground/background classification problem. We use the binary cross-entropy (BCE) loss to supervise the segmentation map at all five scales $\{D_i\}_{i=1}^{5}$. The output based on D_1 is taken as the final prediction.

For referring image localization, we use $\widetilde{D}_5$ as input to predict the bounding box of the target region. Inspired by the anchor-free method [67], we define the ground-truth bounding box as $B = \{x_1, y_1, x_2, y_2\}$, where (x_1, y_1) and (x_2, y_2) are the coordinates of the left-top and right-bottom corners, respectively. We define a 4-D distance vector $\boldsymbol{r} = (l, t, r, b)$, that is, if location (x, y) falls into the ground-truth box, the distances from it to the left boundary, top boundary, right boundary, and bottom boundary to the location (x, y) are computed as

$$l = x - x_1,\ \ t = y - y_1,\ \ r = x_2 - x,\ \ b = y_2 - y. \tag{10}$$

Thus, we use the box-regression branch to predict a 4-D vector (l', t', r', b') to produce the bounding box. This vector is supervised by the intersection over union (IoU) loss. In addition, following [67], we also use a centerness branch, where the centerness of a location is defined as

$$\mathrm{centerness} = \sqrt{\frac{\min(l,r)}{\max(l,r)}\times\frac{\min(t,b)}{\max(t,b)}}. \tag{11}$$

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
QUANTITATIVE RESULTS OF PREC@0.5 ON THE REFERRING IMAGE LOCALIZATION TASK. "-" DENOTES NO AVAILABLE RESULTS

*	UNC			UNC+			G-Ref	
	val	testA	testB	val	testA	testB	val	test
CMN_{16} [62]	-	71.03	65.77	-	54.32	47.76	-	-
MMI_{16} [2]	-	64.90	54.51	-	54.03	42.81	-	-
CG_{17} [3]	-	67.94	55.18	-	57.05	43.33	-	-
$Attr_{17}$ [4]	-	72.08	57.29	-	57.97	46.20	-	-
CMN_{17} [63]	-	71.03	65.77	-	54.32	47.76	-	-
VC_{18} [64]	-	73.33	67.44	-	58.40	53.18	-	-
$ParalAttn_{18}$ [8]	-	75.31	65.52	-	61.34	50.86	-	-
$MAttNet_{18}$ [6]	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
$FAOA_{19}$ [11]	71.15	74.88	66.32	56.86	61.89	49.46	59.44	58.90
$NMTree_{19}$ [8]	71.65	74.81	67.34	58.00	61.09	53.45	61.01	61.46
$lang2seg_{19}$ [23]	77.08	80.34	70.62	-	-	-	65.83	65.44
$RCCF_{20}$ [13]	-	81.06	71.85	-	70.35	56.32	-	65.73
MCN_{20} [25]	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
$Ours_{vanilla}$	80.96	83.56	76.49	68.09	72.97	58.97	71.20	70.89
$Ours_{asy}$	81.09	83.86	75.90	68.56	73.61	59.46	74.33	73.85

TABLE II
QUANTITATIVE RESULTS OF OVERALL IOU ON THE REFERRING IMAGE SEGMENTATION TASK. "-" DENOTES NO AVAILABLE RESULTS

*	ReferIt	UNC			UNC+			G-Ref
	test	val	testA	testB	val	testA	testB	val*
$LSTM\text{-}CNN_{16}$ [16]	48.03	-	-	-	-	-	-	28.14
$RMI+DCRF_{17}$ [17]	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
DMN_{18} [19]	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
KWA_{18} [21]	59.19	-	-	-	-	-	-	36.92
$RRN+DCRF_{18}$ [18]	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
$MAttNet_{18}$ [6]	-	56.51	62.37	51.70	46.67	52.39	40.08	-
$lang2seg_{19}$ [23]	-	58.90	61.77	53.81	-	-	-	-
$CMSA+DCRF_{19}$ [20]	63.80	58.32	60.61	55.09	43.76	47.60	37.89	39.98
$STEP_{19}$ [22]	64.13	60.04	63.46	57.97	48.19	52.33	40.41	46.40
$BRINet+DCRF_{20}$ [1]	63.46	61.35	63.37	59.57	48.57	52.87	42.13	48.04
$LSCM+DCRF_{20}$ [65]	66.57	61.47	64.99	59.55	49.34	53.12	43.50	48.05
$CMPC+DCRF_{20}$ [66]	65.53	61.36	64.54	59.64	49.56	53.44	43.23	49.05
MCN_{20} [25]	-	62.44	64.20	59.71	50.62	54.99	44.69	-
$Ours_{vanilla}$	68.28	64.47	67.22	61.68	51.81	54.77	44.35	52.02
$Ours_{asy}$	68.35	64.40	67.39	61.07	52.13	55.58	44.54	52.66
$Ours_{vanilla}$+DCRF	68.39	64.68	67.33	61.79	51.93	54.85	44.61	55.11
$Ours_{asy}$+DCRF	68.50	64.67	67.46	61.16	52.21	55.63	44.72	52.73

If the location is closer to the center of the object, the centerness is closer to 1. Otherwise, the centerness is closer to 0. This branch is supervised by the BCE loss. Last, we use the focal loss [14] to supervise the classification branch in Fig. 1. Since referring image localization is a single-target prediction task, we directly select the bounding box with the highest score in the classification branch as the output during testing.

IV. EXPERIMENTS AND RESULTS

A. Datasets

We evaluate our method on four datasets, including UNC [68], UNC+ [68], Google-Ref [2], and ReferIt [69].

1) UNC: The UNC dataset is composed of 19 994 images with 142 209 sentences for 50 000 referred objects. It is selected from the MS COCO [70] dataset through a two-player game [69]. The referring expression contains location and appearance information.

2) UNC+: The UNC+ dataset is similar to the UNC dataset, which contains 141 564 sentences to 49 856 targets in 19 992 images. It is also collected from MS COCO [70] with the same strategy. However, there are no location-describing words in its referring expressions. Therefore, we need to understand the appearance and scene context information to locate the target.

3) Google-Ref: Google-Ref has 26 711 images with 104 560 sentences for 54 822 objects. All images and ground-truth masks are still selected from the MS COCO [70] and language descriptions have come from Mechanical Turk. The average sentence length (8.43 words) of Google-Ref is longer than other datasets. In particular, this dataset contains two splits [23]. The first split randomly partitions objects into training and validation sets. The validation set is denoted as "val*." The second split randomly partitions images into training, validation, and testing sets, and the validation and testing sets are denoted as "val" and "test," respectively.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

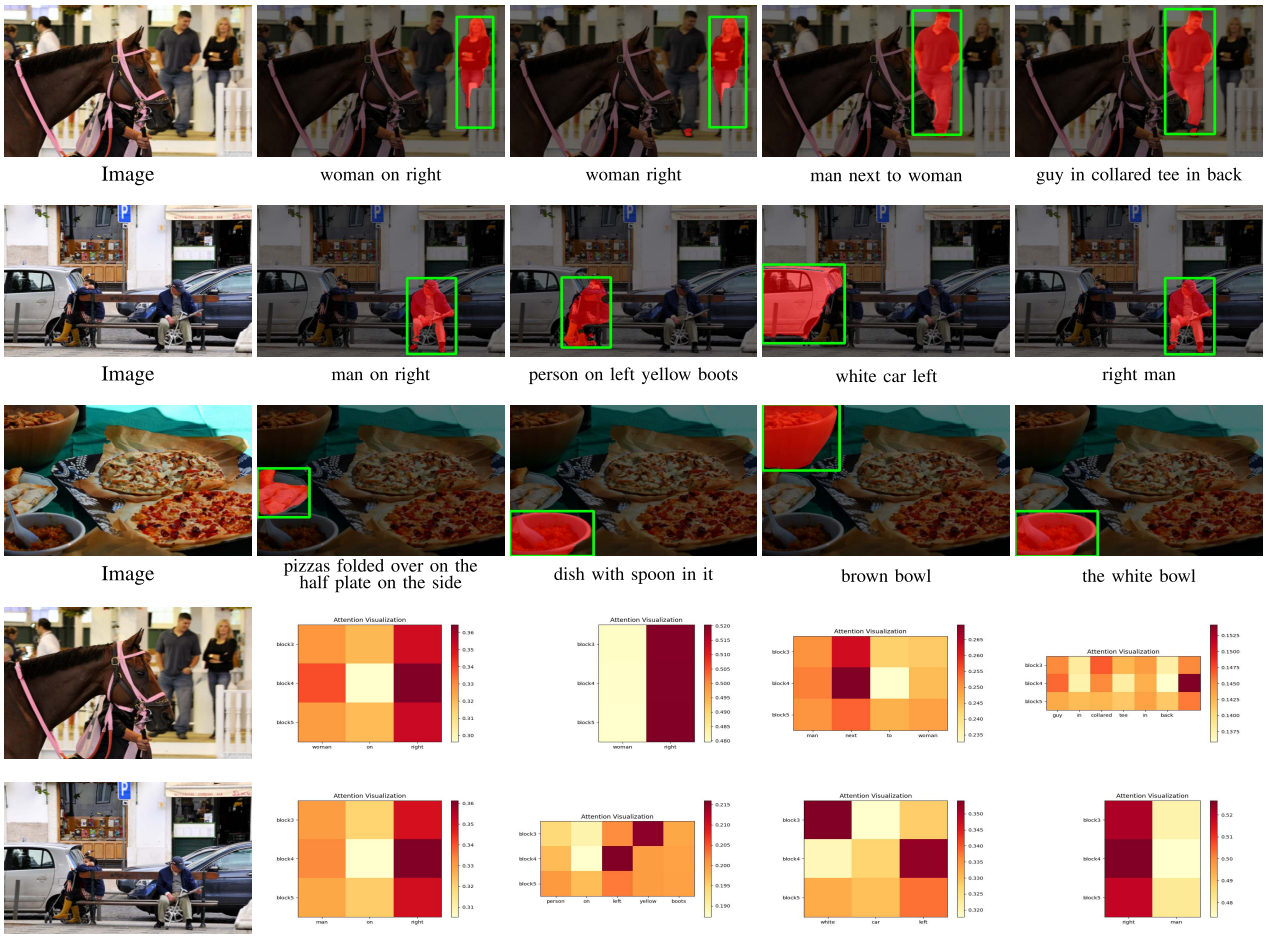

Fig. 3. Qualitative cases of referring image localization and referring image segmentation.

TABLE III
SEGMENTATION RESULTS (IOU) WITH DIFFERENT LENGTH EXPRESSIONS ON UNC, UNC+, GOOGLE-REF, AND REFERITGAME

	Length	1-2	3	4-5	6-20
UNC	R+LSTM [17]	43.66	40.60	33.98	24.91
	R+RMI [17]	44.51	41.86	35.05	25.95
	$\text{Ours}_{\text{vanilla}}$	71.80	67.72	60.75	47.46
	Ours_{asy}	71.54	67.06	60.92	48.37
	Length	1-2	3	4-5	6-20
UNC+	R+LSTM [17]	34.40	24.04	19.31	12.30
	R+RMI [17]	35.72	25.41	21.73	14.37
	$\text{Ours}_{\text{vanilla}}$	63.29	52.39	45.57	34.89
	Ours_{asy}	64.02	52.37	45.85	35.37
	Length	1-5	6-7	8-10	11-20
G-Ref	R+LSTM [17]	32.29	28.27	27.33	26.61
	R+RMI [17]	35.34	31.76	30.66	30.56
	$\text{Ours}_{\text{vanilla}}$	60.28	53.08	50.39	46.71
	Ours_{asy}	60.11	53.38	50.06	48.99
	Length	1	2	3-4	5-20
ReferIt	R+LSTM [17]	67.64	52.26	44.87	33.81
	R+RMI [17]	68.11	52.73	45.69	34.53
	$\text{Ours}_{\text{vanilla}}$	81.19	68.40	61.47	49.01
	Ours_{asy}	81.16	68.61	61.71	49.07

4) ReferIt: ReferIt comprises 19 894 images with 130 525 sentences for 96 654 objects. This dataset is built upon the IAPR TC-12 [71] dataset. In particular, its ground-truth masks contain object and stuff (e.g., water and grass) classes.

B. Implementation Details

Following previous work [25], we resize the input image to 416×416. Darknet-53 [15] is used as visual feature extractor, and its pretraining parameters come from Yolact [40]. The size of word embedding e_t reduces to 512-D. The maximum length of referring expression is restricted to 20. This is because most of the language expressions on the benchmark datasets are shorter than the predefined maximum length, which ensures the integrity of the input sentence in most cases. Our network is built on the public PyTorch toolbox and is trained on two Nvidia GTX 1080Ti GPU for 100 000 iterations. The batch size is set to 16. It is trained with Adam [72] optimizer, and the initial learning rate and weight decay are set to $1e^{-}5$ and $5e^{-}4$, respectively. After 60 000 iterations, the initial learning rate is divided by 10. Also, when training G-ref, we use the UNC model as a pretraining model to avoid overfitting. During the inference phase, the prediction map is resized to the same resolution as the original image.

Metrics: For referring image localization and segmentation, we employ two metrics to evaluate the accuracy: overall IoU and Precision@X. The overall IoU metric calculates the ratio of the total intersection regions and the total union regions between the prediction box/mask and the ground truth.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
RUNTIME ANALYSIS OF DIFFERENT METHODS. THE TIME OF POSTPROCESSING IS IGNORED

	LSTM	RMI	RRN	CMSA	BRINet	MCN	CMPC	$\text{Ours}_{\text{vanilla}}$	Ours_{asy}
Speed	18 FPS	14 FPS	24 FPS	13 FPS	9 FPS	18 FPS	17 FPS	59 FPS	57 FPS

TABLE V
ABLATION STUDY OF REFERRING IMAGE LOCALIZATION ON THE UNC DATASET

	DFS	EFS	VLAM	V-LVAM	A-LVAM	SBAM	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	overall IoU
val	✓						74.83	72.06	66.94	57.90	32.83	62.29
		✓					78.44	75.43	71.28	62.78	39.62	66.96
		✓	✓				78.77	75.71	71.29	62.27	39.07	66.67
		✓	✓	✓			80.27	77.02	72.23	62.35	37.64	67.62
		✓	✓		✓		80.36	77.24	72.46	62.86	37.84	67.70
		✓	✓	✓		✓	80.96	77.86	73.04	64.14	40.17	68.64
		✓	✓		✓	✓	81.09	77.88	73.11	64.05	40.18	68.60
testA	✓						78.35	75.53	71.19	62.10	35.94	64.81
		✓					81.14	78.13	74.21	66.70	42.51	69.31
		✓	✓				82.57	79.49	75.80	67.67	42.41	70.31
		✓	✓	✓			83.84	80.59	76.05	67.42	41.45	70.93
		✓	✓		✓		83.14	80.33	76.45	67.01	40.37	70.44
		✓	✓	✓		✓	83.56	80.33	76.51	68.08	44.00	71.38
		✓	✓		✓	✓	83.86	80.96	76.52	68.45	43.45	71.30
testB	✓						71.60	67.69	61.63	51.36	26.58	58.73
		✓					74.25	70.57	64.48	56.04	34.07	61.71
		✓	✓				75.27	71.52	65.48	56.43	33.95	62.91
		✓	✓	✓			76.04	71.70	65.59	56.57	33.44	62.92
		✓	✓		✓		75.49	71.19	64.71	55.88	32.70	62.66
		✓	✓	✓		✓	76.49	72.03	65.95	57.21	35.70	64.05
		✓	✓		✓	✓	75.90	71.60	65.69	57.06	35.49	63.20

TABLE VI
ABLATION STUDY OF REFERRING IMAGE SEGMENTATION ON THE UNC DATASET

	DFS	EFS	VLAM	V-LVAM	A-LVAM	SBAN	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	overall IoU
val	✓						65.33	60.01	53.34	42.27	18.27	57.99
		✓					70.85	67.45	61.96	50.62	23.43	61.73
		✓	✓				73.41	69.49	63.69	51.56	24.23	62.78
		✓	✓	✓			76.77	72.88	66.65	54.29	25.17	64.44
		✓	✓		✓		76.36	72.21	65.98	53.54	24.85	64.27
		✓	✓	✓		✓	76.79	72.44	65.85	53.19	23.34	64.47
		✓	✓		✓	✓	76.73	72.23	65.95	53.11	22.96	64.40
testA	✓						67.62	62.67	55.01	43.89	17.02	60.11
		✓					74.81	70.81	64.91	53.95	21.94	64.59
		✓	✓				76.75	72.97	66.93	55.68	23.07	65.61
		✓	✓	✓			79.76	76.15	69.77	57.96	24.66	67.53
		✓	✓		✓		79.42	75.46	68.94	57.27	23.88	67.21
		✓	✓	✓		✓	79.48	75.41	69.21	57.08	22.47	67.22
		✓	✓		✓	✓	79.53	75.76	68.53	56.76	22.40	67.39
testB	✓						61.57	55.86	49.44	39.21	19.73	54.62
		✓					65.81	61.37	55.84	46.10	24.44	57.68
		✓	✓				68.01	63.67	57.80	47.44	25.53	59.31
		✓	✓	✓			72.21	67.42	60.84	50.54	26.20	61.04
		✓	✓		✓		72.31	67.32	61.04	50.27	26.24	60.62
		✓	✓	✓		✓	71.82	67.18	60.41	49.56	26.10	61.68
		✓	✓		✓	✓	71.78	66.67	60.47	48.58	25.67	61.07

The second metric calculates the percentage of the images with IoU higher than the threshold X during the testing process, where $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$.

C. Performance Comparison

1) Quantitative Evaluation: The performance (prec@0.5 or IoU) of our model with other existing state-of-the-art

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

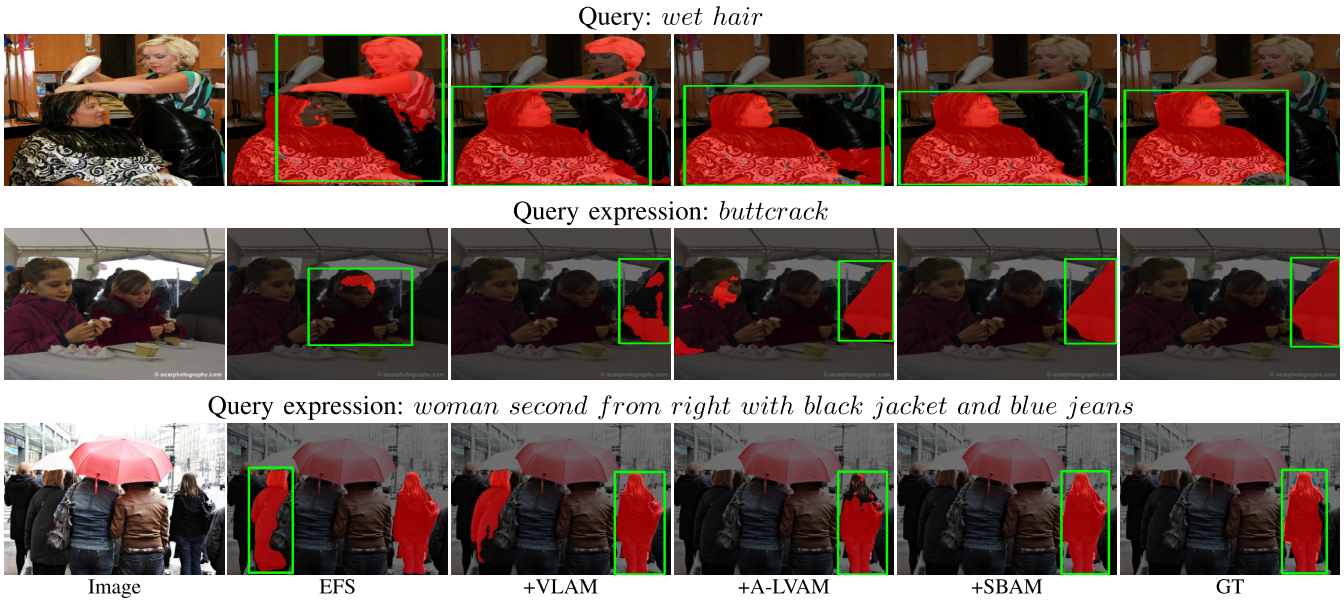

Fig. 4. Visual results generated by different modules.

methods on all datasets is shown in Tables I and II, in which $\text{Ours}_{\text{vanilla}}$ and Ours_{asy} represent the results of using vanilla language-guided visual attention and asymmetric language-guided visual attention, respectively. Our proposed model outperforms the other methods on both localization and segmentation. For the G-Ref, it has longer referring expressions and more complex semantics. Our method leads to a 7.4% improvement in the segmentation task. For the localization, we achieve the gain of 11.8% and 11.9% over the second best method MCN [25] on the G-Ref val and G-Ref test, respectively. These results prove that our method can better capture the long-range dependencies between cross-modal features. The referring expressions of the UNC+ only contain the appearance and scene context information, and they lack the description of the target location. Therefore, the performance enhancement on the UNC+ can show that our method can understand various forms of referring expressions without preference for a specific type of information. Following the work [17], we analyze the effect of sentence length on segmentation performance. The sentences are divided into four groups, and the segmentation results of each group are shown in Table III. Obviously, our method achieves the best performance at any length. We report the running time of different methods in Table IV. The speed of $\text{Ours}_{\text{vanilla}}$ and Ours_{asy} is 59 and 57 FPS, respectively. Our network adopts EFS, which introduces language into the encoder network in a residual manner. The whole network no longer needs additional multimodal feature interaction process. Previous methods first adopt CNN/RNN encoder to extract visual/linguistic features, then use feature interaction layers to realize the cross-modal information fusion, and finally decode the fused multimodal features. The intermediate interaction processing increases the depth of the network, thereby affecting the speed of the model.

2) Qualitative Evaluation: We show some visual results of localization and segmentation in Fig. 3. These results show that our method can accurately identify the objects of interest from the referring expressions of various lengths. For complex multiobjective scene, our method can still do well. Moreover, in the absence of location described in the referring expression, our method can still successfully accomplish the localization and segmentation tasks with the help of appearance attributes (row 3). In the fourth and fifth rows, we give the attention distributions of the linguistic representation at different levels. Our model can adaptively learn useful language information, thereby facilitating the language-to-vision relationship modeling.

D. Ablation Study

We perform ablation experiments on the UNC dataset to verify the effectiveness of each component. The results are presented in Tables V and VI.

1) Comparison of Encoder Fusion and Decoder Fusion: We first remove BCAM and SBAM from the complete network (Fig. 1) to construct the baseline structure. Then, we use v_i in (1) as the fused multimodal feature, which only contains the final state of Bi-GRU. Next, we compare the performance of introducing v_i into encoder or decoder (traditional methods [13], [20]). The results of referring localization and segmentation are shown in Tables V and VI, respectively. We can find that the EFS is significantly better than decoder fusion strategy (DFS).

2) Effectiveness of VLAM: The VLAM can learn the keywords of each image region. We further introduce VLAM into baseline EFS. For the segmentation task, the VLAM brings 1.7%, 1.6%, and 2.8% IoU improvement on UNC-val, UNC-testA, and UNC-testB, respectively. Similarly, the VLAM consistently improves the performance on all tested localization datasets. Experimental comparisons verify that the adaptive linguistic features are useful.

3) Effectiveness of V-LVAM and A-LVAM: Tables V and VI present the ablation results of V-LVAM and A-LVAM, respectively. We can find that V-LVAM and A-LVAM have similar

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
ANALYSIS OF THE SUFFICIENCY AND EFFECTIVENESS OF SPATIAL COORDINATES

*	Referring Image Localization			Referring Image Segmentation		
	UNC-val	UNC-testA	UNC-testB	UNC-val	UNC-testA	UNC-testB
BRINet w/o SP	80.02	82.91	75.51	64.00	66.59	60.39
BRINet w/ 8-D SP	81.09	83.86	75.90	64.40	67.39	61.07
BRINet w/ 12-D SP	80.54	83.52	75.90	64.38	67.18	61.01

TABLE VIII
ANALYSIS OF THE EFFECTIVENESS OF BCAM AND h_T

*	Referring Image Localization			Referring Image Segmentation		
	UNC-val	UNC-testA	UNC-testB	UNC-val	UNC-testA	UNC-testB
Our_{asy}	81.09	83.86	75.90	64.40	67.39	61.07
Our_{CMSA}	79.08	81.36	73.26	61.98	65.02	58.19
Our_{asy} w/o h_T	80.38	82.96	75.13	63.93	66.86	60.82

TABLE IX
RESULTS OF REFERRING IMAGE LOCALIZATION AND SEGMENTATION WITH DIFFERENT INPUT RESOLUTIONS

*	Referring Image Localization			Referring Image Segmentation		
	UNC-val	UNC-testA	UNC-testB	UNC-val	UNC-testA	UNC-testB
Our_{asy}(320×320)	80.04	82.92	75.13	63.40	66.25	60.14
Our_{asy}(416×416)	81.09	83.86	75.90	64.40	67.39	61.07
Our_{asy}(480×480)	81.26	83.98	76.37	64.55	68.01	61.38

performance, and they show superiority in both localization and segmentation tasks. These results indicate that the guidance of language to vision promotes the cross-modal alignment. Especially in the pixel-level segmentation task, it is more important for language to guide the visual context.

4) Effectiveness of SBAM: The gated bottom-up feature augmentation module plays a vital role in controlling multilevel message passing and fusion. We can find that the localization results after SBAM show better performance.

5) Effectiveness of Spatial Coordinates: We verify the effectiveness of the spatial coordinates on the asymmetric version of BRINet. Specifically, we analyze BRINet without spatial features (BRINet w/o SP), BRINet with the 8-D spatial features, and BRINet with the 12-D spatial features (including the extra coordinates of the top-right corner and bottom-right corner). The experimental results on the UNC dataset are shown in Table VII. We can find that BRINet w/ 8-D SP achieves the best performance, which shows that 8-D features have provided sufficient spatial information.

6) Effectiveness of Pretiled Global Semantic h_T: In (1), the tiled language feature h_T depicts the global semantic cues of language expression. We compare the performance of w/o h_T and w/ h_T in Table VIII, which indicates that the global cues can promote the cross-modal fusion.

7) Compared With Other Cross-Modal Fusion Modules: We replace BCAM with the CMSA module [20] for cross-modal fusion and then retrain the entire network. The results (Our_{CMSA}) are shown in Table VIII. We can see that BCAM shows obvious performance advantages.

8) Results With Different Resolutions: Table IX shows the results of referring image localization and segmentation with different input resolutions. We find that larger input resolution will improve the performance by providing richer information.

9) Visual Cases: In Fig. 4, we visualize some cases to illustrate the function of each component. BCAM can promote a higher consistency in the representation of multimodal information in the feature space, thereby realizing the mutual embedding of linguistic and visual features. SBAM can adaptively select useful information in the process of cross-level feature fusion, which promotes more accurate localization.

10) Computation and Memory Statistics: We quantitatively evaluate V-LVAM, A-LVAM, vanilla BCAM (BCAM_{van}), and asymmetric BCAM (BCAM_{asy}) in Table X. For the V-LVAM/ BCAM_{van}, it can be found that large input size leads to a sharp increase in computational cost and memory usage. This is because the size of the pixel-level affinity matrix will increase exponentially with the increase of image size. In contrast, A-LVAM/BCAM_{asy} utilizes the pixel-level adaptive linguistic features and pooled features of each visual subregion to learn an asymmetric affinity matrix, which greatly reduces the size of the matrix.

E. Failure Cases

Fig. 5 presents some interesting failure cases. One type of failure occurs when the boundary is ambiguous (the first row). This phenomenon can be alleviated by introducing a contrast enhancement mechanism to capture more accurate boundary.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Fig. 5. Visual examples of the failure cases.

TABLE X
GPU MEMORY AND FLOPS COMPARISON. THE LOWER VALUES ARE THE BETTER

Input size	Method	FLOPs(G)	Memory(MB)
512×26×26	V-LVAM	1.3	18.90
	BCAM_{van}	1.48	20.59
	A-LVAM	1.06	14.33
	BCAM_{asy}	1.25	15.61
512×52×52	V-LVAM	8.01	89.43
	BCAM_{van}	8.72	90.67
	A-LVAM	4.26	35.00
	BCAM_{asy}	4.98	40.24
512×104×104	V-LVAM	77.05	1009.59
	BCAM_{van}	79.89	1031.54
	A-LVAM	17.04	120.28
	BCAM_{asy}	19.92	142.24
512×208×208	V-LVAM	1028.38	14710.18
	BCAM_{van}	1039.74	14794.64
	A-LVAM	68.16	449.26
	BCAM_{asy}	79.70	533.73

Another case is that when the query is ambiguity and difficult to understand (the second row). Finally, we find that some queries contain misspelled words (nana → banana in the third row), which will lead to the loss of part of the semantic information in the sentence and produce incorrect prediction results.

V. CONCLUSION

We have proposed a BRINet for referring image localization and segmentation. First, a BCAM is used to model the dependence between language and vision. BCAM contains the VLAM and the LVAM. Together they encourage accurate and consistent semantic representations between cross-modal features, thereby realizing the cross-modal mutual guidance. Second, a SBAM employs the segmentation map to control the message passing among the multilevel features, which makes multilevel clues better integrated. The experimental results on UNC, UNC+, Google-Ref, and ReferIt witness the performance gains of referring image localization and referring image segmentation.

REFERENCES

[1] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4424–4433.
[2] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 11–20.
[3] R. Luo and G. Shakhnarovich, "Comprehension-guided referring expressions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7102–7111.
[4] J. Liu, L. Wang, and M.-H. Yang, "Referring expression generation and comprehension via attributes," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4856–4864.
[5] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7746–7755.
[6] L. Yu *et al.*, "MAttNet: Modular attention network for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1307–1315.
[7] S. Yang, G. Li, and Y. Yu, "Cross-modal relationship inference for grounding referring expressions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4145–4154.
[8] D. Liu, H. Zhang, Z.-J. Zha, and F. Wu, "Learning to assemble neural module tree networks for visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4673–4682.
[9] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4644–4653.
[10] X. Liu, L. Li, S. Wang, Z.-J. Zha, D. Meng, and Q. Huang, "Adaptive reconstruction network for weakly supervised referring expression grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2611–2620.
[11] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4683–4693.
[12] A. Sadhu, K. Chen, and R. Nevatia, "Zero-shot grounding of objects from natural language queries," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4694–4703.
[13] Y. Liao *et al.*, "A real-time cross-modality correlation filtering method for referring expression comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10880–10889.
[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[16] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 108–124.
[17] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1271–1280.
[18] R. Li *et al.*, "Referring image segmentation via recurrent refinement networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5745–5753.
[19] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Proc. ECCV*, 2018, pp. 630–645.
[20] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10502–10511.
[21] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 38–54.
[22] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, and T.-L. Liu, "See-through-text grouping for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7454–7463.
[23] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang, "Referring expression object segmentation with caption-aware consistency," in *Proc. BMVC*, 2019, pp. 1–12.
[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[25] G. Luo *et al.*, "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10034–10043.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[30] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. ICCV*, 2015, pp. 1520–1528.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[32] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2393–2402.

[33] D. Lin, Y. Ji, D. Lischinski, D. Cohen-Or, and H. Huang, "Multi-scale context intertwining for semantic segmentation," in *Proc. ECCV*, 2018, pp. 603–619.

[34] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. ACCV*. Cham, Switzerland: Springer, 2016, pp. 213–228.

[35] X. Chen *et al.*, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," 2020, *arXiv:2007.09183*. [Online]. Available: http://arxiv.org/abs/2007.09183

[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[39] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4974–4983.

[40] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.

[41] J. Yu, J. Yao, J. Zhang, Z. Yu, and D. Tao, "Single pixel reconstruction for one-stage instance segmentation," 2019, *arXiv:1904.07426*. [Online]. Available: http://arxiv.org/abs/1904.07426

[42] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 282–298.

[43] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 649–665.

[44] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOV2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–17.

[45] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4555–4564.

[46] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, and A. V. D. Hengel, "Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1960–1968.

[47] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[48] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[49] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[50] Y. Lu, J. Lu, S. Zhang, and P. Hall, "Traffic signal detection and classification in street views using an attention model," *Comput. Vis. Media*, vol. 4, no. 3, pp. 253–266, Sep. 2018.

[51] H. Zhao *et al.*, "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. ECCV*, 2018, pp. 267–283.

[52] B. Chen, P. Li, C. Sun, D. Wang, G. Yang, and H. Lu, "Multi attention module for visual tracking," *Pattern Recognit.*, vol. 87, pp. 80–93, Mar. 2019.

[53] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "DiSAN: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. AAAI*, 2018, pp. 1–10.

[54] L. Li, S. Wang, S. Jiang, and Q. Huang, "Attentive recurrent neural network for weak-supervised multi-label image classification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1092–1100.

[55] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8838–8848.

[56] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 3534–3545, Jan. 2020.

[57] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15506–15515.

[58] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054384.

[59] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 13, 2021, doi: 10.1109/TPAMI.2021.3079993.

[60] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10313–10322.

[61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[62] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 817–834.

[63] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1115–1124.

[64] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4158–4166.

[65] T. Hui *et al.*, "Linguistic structure guided context modeling for referring image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 59–75.

[66] S. Huang *et al.*, "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10488–10497.

[67] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[68] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. ECCV*. Cham, Switzerland: Springer, 2016, pp. 69–85.

[69] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 787–798.

[70] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 740–755.

[71] H. J. Escalante *et al.*, "The segmented and annotated IAPR TC-12 benchmark," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 419–428, 2010.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Guang Feng received the B.E. degree in electronic information engineering from Qingdao University, Qingdao, China, in 2015, and the M.E. degree in signal and information processing from the University of Jinan, Jinan, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China.

His research interests include saliency detection and referring expression comprehension.

Zhiwei Hu received the M.Eng. degree in electronics and communication engineering from Dalian University of Technology (DUT), Dalian, China, in 2021.

His research interests include computer vision and natural language processing.

Lihe Zhang (Member, IEEE) received the M.S. degree from Harbin Engineering University (HEU), Harbin, China, in 2001, and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004.

He is currently a Full Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT), Dalian, China. His current research interests include computer vision and pattern recognition.

Jiayu Sun received the B.E. degree in biomedical engineering from Northeastern University, Shenyang, China, in 2016, and the M.E. degree in electrical engineering from The University of Queensland, Brisbane, QLD, Australia, in 2018. She is currently pursuing the Ph.D. degree with Dalian University of Technology, Dalian, China.

Her research interests include computer vision and machine learning.

Huchuan Lu (Senior Member, IEEE) received the M.S. degree in signal and information processing and the Ph.D. degree in system engineering from Dalian University of Technology (DUT), Dalian, China, in 1998 and 2008, respectively.

He is currently a Full Professor with the School of Information and Communication Engineering, DUT. His current research interests include computer vision and pattern recognition with a focus on visual tracking, saliency detection, and segmentation.

Dr. Lu is also an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Authorized licensed use limited to: Dalian University of Technology. Downloaded on March 30,2022 at 08:27:53 UTC from IEEE Xplore. Restrictions apply.