Referring Expression Object Segmentation with Caption-Aware Consistency

Yi-Wen Chen[1]
chenyiwena@gmail.com

Yi-Hsuan Tsai[2]
ytsai@nec-labs.com

Tiantian Wang[3]
tiantianwang.ice@gmail.com

Yen-Yu Lin[1]
yylin@citi.sinica.edu.tw

Ming-Hsuan Yang[34]
mhyang@ucmerced.edu

[1] Academia Sinica

[2] NEC Laboratories America

[3] University of California, Merced

[4] Google Cloud

Abstract

Referring expressions are natural language descriptions that identify a particular object within a scene and are widely used in our daily conversations. In this work, we focus on segmenting the object in an image specified by a referring expression. To this end, we propose an end-to-end trainable comprehension network that consists of the language and visual encoders to extract feature representations from both domains. We introduce the spatial-aware dynamic filters to transfer knowledge from text to image, and effectively capture the spatial information of the specified object. To better communicate between the language and visual modules, we employ a caption generation network that takes features shared across both domains as input, and improves both representations via a consistency that enforces the generated sentence to be similar to the given referring expression. We evaluate the proposed framework on two referring expression datasets and show that our method performs favorably against the state-of-the-art algorithms.

1 Introduction

Object segmentation aims to separate foreground objects from the background. In this work, we focus on object segmentation from referring expressions, in which the segmentation is guided by a natural language description that identifies a particular object instance in a scene, *e.g.*, *the man in a blue jacket* or *the laptop on the left*.

Transferring knowledge between language and visual domains is an important but challenging task. Two relevant tasks are: 1) *referring expression comprehension* for localizing or segmenting an object according to a natural language description, and 2) *referring expression generation* for producing a sentence that identifies a particular object in an image. Existing methods [15, 24] address both tasks by constructing a generation model and inferring the region which maximizes the expression posterior in the comprehension task. However, such joint information is usually exploited to only enhance the generation performance.
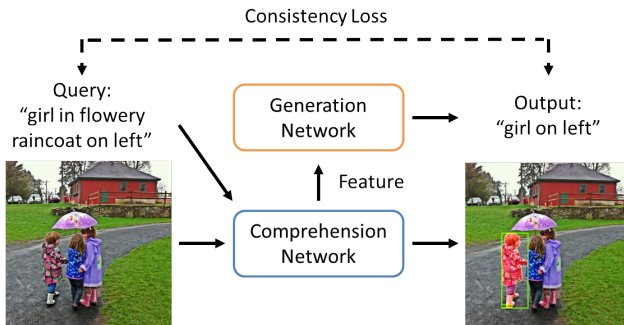
Figure 1: Illustration of the proposed algorithm. Given an image and a referring expression, we use a comprehension network to segment the specified object. With the features containing both language and visual information as input, the generation network produces a sentence identifying the target object. By enforcing a caption-aware consistency loss between the query and output sentence, we further improve the performance of comprehension.

In this paper, we focus on referring expression object segmentation. Unlike existing methods, our model jointly considers both tasks to benefit the comprehension task. Intuitively, when one signal, *e.g.*, a sentence, is transferred from the language domain to the visual domain, and then transferred back to the language domain, the transferred-back signal is supposed to be similar to the original one. By exploiting this property, we develop a network that jointly considers referring expression comprehension and generation, and enforces a caption-aware consistency between the visual and language domains.

To this end, we first design a comprehension network that contains the language and visual encoders to extract the feature representations of respective domains. To connect these two domains, we further propose to use *spatial-aware dynamic filters* to bridge the language and visual encoders. Meanwhile, these filters provide visual representations with the localization ability from the input referring expression. Based on the proposed baseline model, we then employ a caption generation model that takes feature representations from the comprehension network as inputs. The generated referring expression should be similar to the original sentence, and we leverage this property as an additional consistency cue to enhance the language and visual representations. The main steps of the proposed model are illustrated in Figure 1.

To evaluate the proposed method, we conduct extensive experiments on the RefCOCO [24] and RefCOCOg [15, 17] datasets. Experimental results show that our model performs favorably against the state-of-the-art methods. In addition, we provide the ablation study to demonstrate the effectiveness of each component in the proposed framework, including the spatial-aware dynamic filters and caption-aware consistency. The main contributions of this work are summarized as follows: 1) We integrate referring expression generation into referring expression comprehension so that the two complementary tasks can benefit each other via enforcing the caption-aware consistency. 2) We develop the spatial-aware dynamic filters that bridge the visual and language domains and facilitate the feature learning process. 3) We design an end-to-end trainable network for referring expression comprehension, achieving the state-of-the-art performance.

2 Related Work

Referring Expression Comprehension. The task of referring expression comprehension aims to localize or segment an object given a natural language description. Existing methods [6, 14, 15] mainly rely on recurrent caption generation models, and select the object with the maximum posterior probability of the expression among all object proposals. By exploring the relationship between the object and its context [17, 24, 27], the target object can be better localized. Recent approaches adopt various learning strategies, such as embedding images and sentences into a common feature space [19, 22], or learning attributes [13] to help differentiate objects of the same category. In addition, Hu *et al*. [7] analyze the inter-object relationships by parsing the sentence into subject, relationship and object parts. To jointly consider the associated factors such as attributes and relationships between objects, Yu *et al*. [26] propose a modular attention network to decompose the expression into subject appearances, locations, and relationships to other objects.

While the aforementioned methods mainly localize an object by a bounding box, algorithms that focus on segmentation [6, 9, 12, 16, 20] usually encode the referring expression through the LSTM network and use a fully convolutional network for foreground/background segmentation by using both the language and visual features. Different from these approaches, our proposal-based model first localizes objects and performs segmentation via learning better feature representations through a referring generation network that considers the caption consistency. We note that the approach in [19] also considers the consistency between the generated sentence and input sentence but does not target at segmenting objects. Furthermore, this approach uses pre-defined and fixed region proposals, in which the visual representations are not updated through the proposals. In contrast, our unified framework is end-to-end trainable while bridging features across visual and language domains.

Referring Expression Generation. The generation task of referring expressions is a special case of image captioning. Rather than describing the whole image, the generated sentence uniquely identifies an object within the image. A referring expression is considered good if one can localize the corresponding object by comprehending this referring expression. Therefore, referring expression comprehension is often employed in the generation task [13, 14, 15] to improve the performance.

CNN-LSTM based models are widely used for image captioning [8, 18, 21, 23]. While a CNN model extracts visual features, an LSTM module produces captions. To address referring expression generation, Mao *et al*. [15] combine the extracted visual features with the location and size of the target object. Furthermore, this method uses a CNN-LSTM model for the comprehension task and jointly trains the generation and comprehension modules. Yu *et al*. [25] further propose a joint speaker-listener-reinforcer model where a reward function is introduced to guide the expression sampling. Their approach jointly trains the generation and comprehension networks, but does not specifically consider the caption consistency as our framework. While the aforementioned methods mainly utilize the comprehension model to generate high-quality sentences, in this work, we focus on the comprehension task and demonstrate that the generation model also facilitates the comprehension performance by enforcing the proposed caption-aware consistency between the visual and language domains.
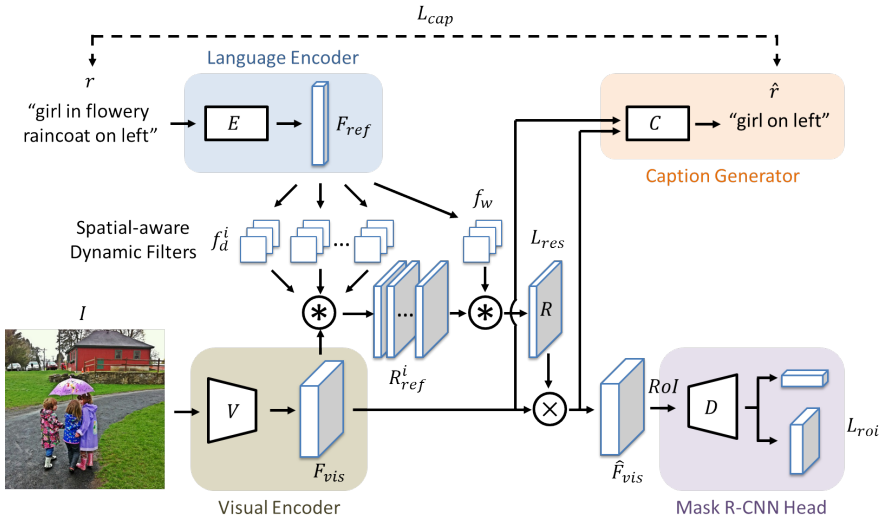
Figure 2: Architecture of the proposed framework. The proposed network is composed of a language encoder E, a visual encoder V, a Mask R-CNN head D, and a caption generator C. Features of the referring expression r and image I are extracted by E and V respectively. Knowledge is transferred from language domain to visual domain via the spatial-aware dynamic filters f_d^i, by which a response map R is generated to produce a location-aware feature $\hat{F}_{vis}$. Based on this feature $\hat{F}_{vis}$ that carries information from both domains, we generate object bounding box and mask by the Mask R-CNN head D. The caption generator takes features F_{vis} and $\hat{F}_{vis}$ as inputs, and produces a sentence identifying the target object. We apply a caption-aware consistency loss between the input query r and the generated sentence $\hat{r}$ to further improve the language and visual feature representations.

3   Proposed Framework

In this work, we focus on referring expression object segmentation. The overview of the proposed framework is illustrated in Figure 2. Given an image I and a natural language description r, we aim to segment the object in I specified by r. To this end, we propose an end-to-end trainable network that contains a language encoder E, a visual encoder V, a Mask R-CNN head D, and a caption generator C. The encoders E and V extract language and visual features, respectively. Motivated by the dynamic filter network [1], we enhance the ability of specific object localization via introducing the *spatial-aware dynamic filters* to transfer knowledge from text to image. The yielded cross-modal information allows the Mask R-CNN head D to produce more accurate segmentation results. To further improve our model, we employ the caption generation network C and a consistency loss L_{cap} to jointly train the comprehension and generation networks. We describe each component of the proposed network below.

3.1   Segmentation from Referring Expression

In this subsection, we introduce how the proposed network generates the object segment given the query referring expression. To this end, the language encoder E, visual encoder V,

and spatial-aware dynamic filters are elaborated.

Language Encoder. Similar to [26], we use a bi-directional LSTM model to extract features of a referring expression. Given a referring expression $r = \{w_t\}_{t=1}^T$ of T words with each word w_t represented by a one-hot vector e_t, the bi-directional LSTM S is applied to encode the whole sentence in both forward and backward directions:

$$\overrightarrow{h}_t = \overrightarrow{S}(e_t, \overrightarrow{h}_{t-1})$$
$$\overleftarrow{h}_t = \overleftarrow{S}(e_t, \overleftarrow{h}_{t+1})$$
$$F_{ref} = [\overrightarrow{h}_T, \overleftarrow{h}_1], \tag{1}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the forward and backward hidden states at time step t, respectively. We concatenate the final hidden states in both directions to yield the feature representation F_{ref} of the referring expression.

Visual Encoder. Given an input image I, we aim at pixel-wise segmentation. Different from the approaches based on the fully convolutional network (FCN) that does not generate instance-aware results, we adopt the proposal-based Mask R-CNN [4] framework to generate an object mask based on each detected object bounding box. We use the ResNet-101 [3] model as the backbone network and extract features over the entire image. The feature from the final convolutional layer of the fourth block, denoted by $F_{vis} = V(I)$, serves as the representation of image I.

Spatial-aware Dynamic Filters. Motivated by the recent work [10] on tracking with natural language, we utilize dynamic convolutional filters as a bridge to connect the language and visual domains. Unlike conventional convolutional filters that apply the same weights to all input images, dynamic convolutional filters are generated depending on the input sentence. Given the feature representation F_{ref} of a sentence r, a single fully connected layer parameterized by the weights W_d^1 and the bias b_d^1 is adopted to generate a set of dynamic filters:

$$f_d^1 = \tanh(W_d^1 \cdot F_{ref} + b_d^1), \tag{2}$$

where tanh is the hyperbolic tangent function, and f_d^1 is a set of 1×1 convolutional filters with the same number of channels as the visual representation F_{vis}. We then convolve the visual representation F_{vis} with the generated dynamic filters f_d^1 to obtain a response map R_{ref}^1:

$$R_{ref}^1 = f_d^1 * F_{vis}. \tag{3}$$

With this formulation, knowledge is transferred from the language domain through learning the dynamic filters, with which the response map reflects the information inferred from the referring expression.

However, such filters consider the entire image and thus may only be able to catch the global structure but ignore spatially distributed objects. As such, we propose to utilize spatial-aware dynamic convolutional filters that consider local regions of the image, including up, down, left, right, horizontal and vertical middle regions, and each region covers a half area of the entire image. We thereby apply six additional fully connected layers to generate spatial-aware dynamic filters $\{f_d^i\}_{i=2}^7$ corresponding to each region i via (2). The six dynamic filters are then convolved with the visual feature F_{vis}, where the values outside the

defined regions are set to 0. Then we obtain six spatial-aware response maps similar to (3), denoted by $\{R^i_{ref}\}^7_{i=2}$, in which each map focuses on its defined region.

To combine these spatial response maps and the one from (3), we adopt another set of dynamic filters f_w with 7 channels, which are also generated from the sentence representation F_{ref}, to account for the importance of each region depending on the input sentence. We convolve f_w with the concatenation of the 7 response maps $R_{con} = \text{concat}(R^i_{ref})$ and obtain a final response map R with one channel, *i.e.*,

$$R = \sigma(f_w * R_{con}), \tag{4}$$

where σ is the sigmoid function with output range $[0, 1]$. Ideally, R represents a map of the object specified by the input referring expression. Thus, we apply a binary cross-entropy loss L_{res} to supervise the response map R with respect to the ground-truth object mask.

Baseline Objective. Based on the response map in (4), we take the element-wise multiplication of R and F_{vis} to be the caption-aware feature representation $\hat{F}_{vis}$, which carries the information from both the language and visual domains. To obtain the final segmentation result, we then feed $\hat{F}_{vis}$ into the Mask R-CNN [4] RoI head D, which includes the bounding box and the binary segmentation branches. The overall objective can be written as:

$$L = L_{roi} + L_{res}, \tag{5}$$

where L_{roi} includes the classification loss, bounding box loss and mask loss, the same as those defined in Mask R-CNN. With this formulation, we construct an end-to-end trainable network that produces the referring expression object segmentation. Unlike the state-of-the-art methods, such as MAttNet [26], that require multiple training stages and pre-processing steps, our model can be efficiently learned, through the help of spatial-aware dynamic filters which provide the spatial information from the input sentence.

3.2 A Joint Framework

In light of the cycle consistency work [28] that solves the domain transfer problem in cross-directions, we integrate both the referring expression comprehension and generation tasks into a joint framework, where their feature representations are shared and can be jointly optimized through back-propagation.

Referring Expression from Segmentation. To generate a sentence describing a particular object within an image, we adopt the attention-based image captioning model [23]. To train the caption generation model C, we input the feature representation F_{vis} extracted from Mask R-CNN and concatenate it with $\hat{F}_{vis}$ which contains the spatial information about the object. As a result, during training the caption generation model, gradients can be back-propagated through F_{vis} to update the Mask R-CNN feature extractor, as well as through $\hat{F}_{vis}$ to optimize dynamic filters and the language encoder.

Caption-aware Consistency. Given the ground-truth sentence $r = \{w_t\}^T_{t=1}$, which is the input to the language encoder, the objective for caption generation is to minimize the cross-entropy loss L_{cap}:

$$L_{cap} = -\sum_{t=1}^{T} \log(p_{\theta_c}(w_t | w_1, ..., w_{t-1})), \tag{6}$$

Table 1: Localization results of our method and the competing methods on two datasets. We summarize the major information used in each method, including context (C), attribute prediction (Attr), attention module (Attn), location (L), relationships between objects (R), and joint training with referring expression generation (J).

Model	Info.	RefCOCO			RefCOCOg		
		val	testA	testB	val*	val	test
Nagaraja *et al.* [■]	C	57.30	58.60	56.40	-	-	49.50
Luo *et al.* [■]	J	-	67.94	55.18	49.07	-	-
Liu *et al.* [■]	Attr, J	-	72.08	57.29	52.35	-	-
Yu *et al.* [■]	J	-	73.78	63.83	59.84	-	-
MAttNet [■]	Attr, Attn, L, R	76.65	81.14	69.99	-	66.58	67.27
VC [■]	C	-	73.33	67.44	62.30	-	-
baseline	-	72.65	76.65	65.75	54.18	58.09	58.32
+ spatial coords [■]	L	75.89	78.57	68.54	61.37	64.10	64.21
+ spatial-aware filters	L	76.98	79.30	69.75	61.65	65.18	65.28
+ caption-aware consistency	J	76.05	78.84	69.36	60.69	64.71	63.79
full model	L, J	**77.08**	**80.34**	**70.62**	**62.34**	**65.83**	**65.44**

where $p_{\theta_c}(\hat{w}_t|\hat{w}_1,...,\hat{w}_{t-1})$ is the probability of predicting a particular word from the caption generation network parameterized by θ_c. Here, this loss function in our framework enforces that the predicted sentence $\hat{r}$ generated by the feature $\hat{F}_{vis}$, *i.e.*, $\hat{r} = C(\hat{F}_{vis}, \cdot)$, should be consistent with the input query r that generates the same feature, *i.e.*, $\hat{F}_{vis} = \mathcal{F}(E(r))$, where $\mathcal{F}$ is a mixed operation involving the visual encoder V and dynamic filters in the proposed method. Hence, our caption-aware consistency actually enforces $r \approx C(\mathcal{F}(E(r)), \cdot)$.

Overall Objective. To exploit the caption-aware consistency, we jointly train the comprehension model, including the language encoder E, visual encoder V, Mask R-CNN head D in Section 3.1 and the caption generation model C in Section 3.2. The total loss function is extended from (5) to:

$$L = L_{roi} + L_{res} + \alpha L_{cap}, \tag{7}$$

where α is the coefficient of the consistency loss. We note that adding L_{cap} enables the joint optimization between the language and the visual domains. That is, the intermediate feature $\hat{F}_{vis}$ would be updated by the guidance from the first two loss functions in (7), which are supervised by the comprehension task, and in the meanwhile L_{cap} updates the feature based on the caption generation task.

3.3 Model Training and Implementation Details

To train the joint network model, we adopt a sequential training strategy to optimize the objective in (7). First, we only update the comprehension network by optimizing (5). Then, we pre-train the caption generation network by optimizing (6) as a warm-up. Finally, we update the entire framework with the objective in (7). With the trained model, we choose the detected object with the largest score during testing.

We implement our model with PyTorch using the SGD optimizer. For the language encoder, the dimension of the LSTM hidden states is set to 512. By concatenating the forward and backward hidden states, the feature F_{ref} is a 1024-dimensional vector. In the visual encoder, the visual feature F_{vis} is of dimension 1024. Thus, we also generate the dynamic filters of dimension 1024. For the caption generation model, the input spatial features are resized to 14×14 and have the same number of channels as that of the concatenation of F_{vis} and $\hat{F}_{vis}$.

Table 2: Segmentation results of our method and the competing methods on two datasets.

		RefCOCO			RefCOCOg		
Model	Backbone Net	val	testA	testB	val*	val	test
D+RMI+DCRF [□]	Deeplab101	45.18	45.69	45.57	-	-	-
RRN+LSTM+DCRF [◨]	Deeplab101	55.33	57.26	53.95	36.45	-	-
MAttNet [26]	Res101	56.51	62.37	51.70	-	47.64	48.61
KWAN [□]	Deeplab101	-	-	-	36.92	-	-
DMN [16]	DPN92	49.78	54.83	45.13	36.76	-	-
Ours	Res101	58.90	61.77	53.81	44.32	46.37	46.95

When training the full model, the loss weight α in (7) is set to 0.1 for all experiments. The codes and models are available at: https://github.com/wenz116/lang2seg.

4 Experimental Results

We evaluate the proposed framework on two referring expression datasets: RefCOCO [24] and RefCOCOg (with two splits[1]) [15, 17]. The two datasets are collected from the Microsoft COCO images [11], with different properties of expressions. We show both detection and segmentation results with comparisons against the state-of-the-art algorithms. In addition, we present an ablation study to demonstrate the importance of each component in the proposed framework. More results are provided in the supplementary material.

For evaluating the detection performance, the predicted bounding box is considered correct if the intersection-over-union (IoU) of the prediction and the ground truth is above 0.5. As for the segmentation quality, we use Intersection-over-Union (IoU) as metric.

4.1 Localization Results

In Table 1, we show comparisons with existing state-of-the-art algorithms [13, 14, 17, 25, 26, 27]. Since each method adopts diverse information to help the comprehension task, we further summarize the major cues that each approach relies on, such as context information [17, 27], attribute prediction [13, 26], and joint training with referring expression generation [13, 14, 25].

Table 1 shows that the proposed method performs favorably against most methods by significant margins, and competitively with MAttNet [26]. We note that, the MAttNet method utilizes various cues, including attention module, attribute prediction, location information, and relations between objects to achieve good performance, while our model only focuses on the location cue and joint training with referring expression generation. It is also worth mentioning that our model is a unified framework that is end-to-end trainable, while MAttNet requires multiple separate training stages to obtain the final model. The runtime speed of our method is 0.17 seconds per image, which is much faster than MAttNet with 0.67 seconds per image on an Intel Xeon 2.5 GHz machine and an NVIDIA GTX 1080 Ti GPU with 11 GB memory.

[1]The first split [15] randomly partitions objects into training and validation sets. We denote the validation set as "val*" in this paper. The second split [17] randomly partitions images into training, validation and testing sets, where we denote the validation and testing ones as "val" and "test", respectively.

Figure 3: Sample results of objects referred by various query expressions.

Figure 4: Sample results from different variants of the proposed model on RefCOCO.

4.2 Segmentation Results

We present the experimental results with comparisons to the state-of-the-art algorithms including D+RMI+DCRF [12], RRN+LSTM+DCRF [9], MAttNet [26], KWAN [20] and DMN [16] on the two datasets in Table 2. Overall, our method consistently and significantly outperforms other segmentation-based approaches that use a similar backbone network (*i.e.*, Deeplab [2] with ResNet-101) as ours. Different from the DMN [16] scheme that utilizes dynamic filters in a sequential manner for capturing the information of each word in a sentence, our model generates the dynamic filters in a spatial-aware manner, where each set of filters produces a response map to certain region of the image. We note that the proposed method achieves better performance. Similar to the localization results, MAttNet [26] that fuses multiple cues performs competitively with our model. We present qualitative examples of referring expression object segmentation in Figure 3. The proposed model can segment different objects according to various query expressions, such as the location, color, or action information, and further demonstrates the effectiveness of the proposed caption-aware consistency framework.

4.3 Ablation Study

We present the results of an ablation study in Table 1. We first show that using the proposed spatial-aware dynamic filters improves the baseline with only a single dynamic filter or the spatial-aware mechanism [5] that concatenates spatial coordinates and feature maps. Second, the referring expression generation network with caption-aware consistency performs favorably against the baseline model. In the full model with both spatial-aware filters and caption-aware consistency, higher performance gains are achieved over other baselines.

We present sample segmentation results predicted by different variants of our model in Figure 4. Compared with the baseline and the model with spatial-aware filters, the proposed full model can localize objects accurately while the baseline model predicts the wrong object. In addition to improving the localizing ability, our full model enhances feature representations around the object. For instance, the elephant in back is well segmented by our model even if it is surrounded by complex background and similar instances.

5   Concluding Remarks

In this paper, we propose an end-to-end trainable framework for referring expression segmentation. We design a comprehension model that consists of language and visual encoders to extract feature representations in the respective domains. By introducing the spatial-aware dynamic filters, knowledge can be transferred from language domain to visual domain, while capturing the useful location cue. In addition to the proposed baseline model, we employ a caption generation network to connect referring expression comprehension and generation. Considering the consistency that the generated sentence is supposed to be similar to the given referring expression, we enforce a caption-aware consistency loss and further enhance the language and visual representations. Extensive experiments and an ablation study on two referring expression datasets show that the proposed algorithm achieves favorable performance against the state-of-the-art methods.

References

[1] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NIPS*, 2016.

[2] 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. *arXiv:1606.00915*, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[5] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016.

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

[7] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017.

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

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

[10] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W.M. Smeulders. Tracking by natural language specification. In *CVPR*, 2017.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

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

[13] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *ICCV*, 2017.

[14] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, 2017.

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

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

[17] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016.

[18] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[19] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016.

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

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[22] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[23] Kelvin Xu, Jimmy L. Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[24] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[25] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017.

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

[27] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *CVPR*, 2018.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.