Referring Image Segmentation by Generative Adversarial Learning

Shuang Qiu, Yao Zhao, *Senior Member, IEEE*, Jianbo Jiao, Yunchao Wei, and Shikui Wei, *Senior Member, IEEE*

***Abstract*—Referring expression is a kind of language expression being used for referring to particular objects. In this paper, we focus on the problem of image segmentation from natural language referring expressions. Existing works tackle this problem by augmenting the convolutional semantic segmentation networks with an LSTM sentence encoder, which is optimized by a pixel-wise classification loss. We argue that the distribution similarity between the inference and ground truth plays an important role in referring image segmentation. Therefore we introduce a complementary loss considering the consistency between the two distributions. To this end, we propose to train the referring image segmentation model in a generative adversarial fashion, which well addresses the distribution similarity problem. In particular, the proposed adversarial semantic guidance network (ASGN) includes the following advantages: a) more detailed visual information is incorporated by the detail enhancement; b) semantic information counteracts the word embedding impact; c) the proposed adversarial learning approach relieves the distribution inconsistencies. Experimental results on four standard datasets show significant improvements over all the compared baseline models, demonstrating the effectiveness of our method.**

***Index Terms*—Image referring segmentation, Adversarial training.**

I. Introduction

ALTHOUGH significant progress has been achieved on semantic image segmentation, the more general and challenging task of segmenting entities based on arbitrary natural language expressions remains far from being solved. In this paper, we study the problem of using natural language expressions to segment an image. Given an image and a natural language expression, we aim at segmenting out the corresponding region referred by the expression, such as "White vehicle on the right" shown in Fig. 1. The expression always contains the attributes and positions of the entities, like colors and relationships. While being a new topic introduced by [1], [2] recently, this problem has great value as it provides a novel approach for interactive image segmentation. For instance, users can segment/select image regions of interest by typing natural language descriptions or directly speaking to the computer [3]. Thus, there is a great impetus to develop effective and accurate tool for the tasks involving photo editing, language-based human-robot interface, and automatic-selected object tracking.

Fig. 1. Given the input images (left) and corresponding referring expressions (below), our model is able to segment out the referred regions (right). Two examples are shown with the ground truth masks in the middle column.

Considering the success of convolutional neural networks (CNNs) in semantic segmentation [4], [5], an intuitive approach to tackle this problem is concatenating the image features produced by the semantic segmentation networks with the sentence representation produced by an LSTM [6] sentence encoder. Such sentence-to-image interaction scheme has been widely adopted by existing methods on referring object localization [7]–[10], referring segmentation [1], [2], caption generation [11], [12] and cross-modal retrieval [13], [14]. However, the output of semantic segmentation is with low-resolution compared to the input image, thus lacks detailed structure information. Without detailed image information, it is hard for the segmentation networks to identify the edges of the target region, which leads to low performance of prediction.

Manuscript received July 26, 2018; revised January 18, 2019 and May 21, 2019; accepted September 4, 2019. Date of publication September 20, 2019; date of current version April 23, 2020. This work was supported in part by National Key Research and Development of China (2016YFB0800404), in part by National Natural Science Foundation of China (61532005 and 61972022), in part by Program of China Scholarships Council (201807095006), and in part by Fundamental Research Funds for the Central Universities (2018JBZ001). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mohammed Daoudi. *(Corresponding author: Yao Zhao.)*

S. Qiu, Y. Zhao, and S. Wei are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: 14120332@bjtu.edu.cn; yzhao@bjtu.edu.cn; shkwei@bjtu.edu.cn).

J. Jiao is with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: jiaojianbo.i@gmail.com).

Y. Wei is with the Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: wychao1987@gmail.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2942480

Aiming at segmenting out the referred regions, referring image segmentation is essentially a pixel-wise classification problem based on pixel-wise classification loss. Despite differences in the network architectures, a common property among pixel-wise loss approaches is that the label of each pixel is predicted independently from each other. However, in practice the labels of pixels in an image are relevant. If we regard each pixel independently, some information of relevance between pixels is likely to lose. For example, some kinds of objects have specific shapes. The predicted masks of humans should be solid and the masks of donuts always contain holes. But the segmentation results of humans and donuts may lose their characteristics when only optimized by pixel-wise classification loss. The context of an image plays an important role when we classify the pixels and improves the performance of these prediction models by reinforcing spatial contiguity. Instead of considering the similarity of pixels independently, it is highly desirable to enforce spatial label contiguity. In prior works, in order to improve the performance of the independent prediction model, various post-processing approaches have been explored. Some of the existing methods build graph structures over the image by Markov Random Field (MRF) [15] or Conditional Random Field (CRF) [16], to capture the context of an image. In [5], post-processing based on CRF on top of the deep network framework has been adopted to refine pixel label predictions. However, this kind of post-processing is rather time-consuming in testing phase which is a restriction for real-time applications.

To address the above-mentioned issues, we propose adversarial semantic guidance network (ASGN) to add more detailed image semantic guidance and introduce a loss term measuring the distribution similarity. Specifically, we concatenate the multi-scale features encoding image details to the combined (with LSTM features) feature space by a skip connection. In addition, as the combination of expression features may introduce semantic ambiguity, the output features of the backbone are connected to the final features before the mask inference, to provide more semantic guidance. On the other hand, besides the pixel-wise classification loss, we introduce the adversarial loss term by adding a discriminator following the mask inference network inspired by the generative adversarial network (GAN) [17] and previous work [18]. In particular, we minimize the difference between the distributions of the network prediction and ground truth by optimizing an objective function. This function combines a conventional cross-entropy loss with an adversarial term. The adversarial term detects the inconsistencies between the two distributions and encourages the referring segmentation model to produce masks that cannot be distinguished from ground-truth ones.

The contributions of this work are summarized as follows:

- We propose adversarial semantic guidance network (ASGN) for referring semantic segmentation, in which the distribution similarity between the network inference and ground truth is measured by a discriminative loss term.
- We leverage multi-scale feature maps from the network backbone to add more detailed visual information for referring segmentation.
- We employ the semantic embedding after introducing the language expressions so as to counteract the word embedding impact.
- Our approach is general and can be embedded in any other state-of-the-art framework for a further improvement. In particular, we achieve competitive results upon the baseline models [1], [2], i.e., 28.06% vs. 36.82% on IoU.

II. Related Work

In this section, we review the most related works to ours in the following three areas: semantic segmentation, referring expression localization, and adversarial learning.

A. Semantic Segmentation

CNNs have have made remarkable progress in the field of image semantic segmentation [19]–[22]. For instance, FCN [4] converted the fully connected layers in VGG network [23] into fully convolutional layers, for pixel-wise dense labeling. However, the output segmentation map was low in resolution–due to the involvement of pooling layers that can increase the receptive field size rapidly after several steps. As a result, a low-resolution label mask was obtained. To address this issue, the mask can be up-sampled using bi-linear interpolation [4], [24]. Other solutions proposed to use dilated convolutions to increase the receptive field size without losing resolution [5], [25]–[27], skip connections to earlier high-resolution layers [4], [24], [28], multi-resolution networks [29]–[31], or depth-adaptive multi-scale convolution layer [32]. DeepLab [5] alleviated this issue by discarding two pooling operations with atrous convolution. With Residual network [5] as its backbone architecture, DeepLab [5], [25], [28] was one of the leading models on Pascal VOC [33]. Similarly, we utilize the DeepLab architecture (with ResNet-101 as backbone) to extract image features in a fully convolutional manner. Some of the existing methods worked on post-processing of segmentation networks, i.e. Markov Random Field (MRF) or Conditional Random Field (CRF) aimed at finding a graph structure over the image [15], [16], [34]–[37]. Different from the post-processing works above, we adopt an adversarial training framework which introduces a discriminative loss as a complementary loss in order to encourage the model to enhance pixel classification accuracy.

B. Generative Adversarial Networks

Generative Adversarial Network (GAN) [17] was first introduced to address the problem of image synthesis with similar quality to real ones. The fundamental idea of GANs was to play a minimax game by training two networks of a generator and a discriminator. The generator tried to produce more realistic image samples from random noise, while the discriminator aimed to distinguish generated images from the real ones. There were also many works that employed GAN under conditions. For example, some works aimed at generating images which were not only indistinguishable from natural images but also matched the constraints from the conditions. Previous works have conditioned GANs on discrete labels, text, and, indeed, images. [38]

first proposed conditional GAN which can generate MNIST digits conditioned on class labels. For conditioning on text format, Reed *et al.* [39] proposed a deep architecture with conditional GAN which can generate realistic images described by the corresponding natural language. Zhang *et al.* [40] proposed stackGAN which can produce images with a larger size than before by given language expression. The image-conditional models have tackled inpainting [41], hole-filling [41], image manipulation guided by user constraints [42], product photo generation [43], domain transfer [44]–[46] and style transfer [47]. Different from previous works, we incorporate the conditional adversarial training scheme into the referring image segmentation task, in which the segmentation network acts as the generator.

C. Referring Expression Grounding

Our work is related to recent work on object grounding with natural language. [48] and [49] used image captioning models [50], [51] to calculate confidence of each region proposal for whether the region contained target object or not. The proposal with the highest confidence was considered as the ground truth result. [7] focused on incorporating better measures of visual context into referring expression models and utilizing the visual difference between objects within an image to help improve performance. [8] used multiple-instance to discover context regions and discover interactions between the object. In [52], by reconstruction with attention mechanism, the correspondence between the description and image region was learned. These works aimed at grounding the target objects instead of segmenting them out. The most relevant works to ours are [1] and [2], which studied the same problem of image segmentation based on referring expressions. Different from the previous approaches, we propose ASGN model to enforce the distribution similarity between the inference and ground truth masks and detect mismatches to refine pixel label predictions simultaneously as well as add more image detailed semantic guidance.

III. Model

In this section, we first give an overview of the proposed ASGN approach. Then we explain the proposed segmentation network in detail. Finally, we present the generative adversarial framework and introduce the loss function of the proposed method.

A. Overview

In the task of referring image segmentation, an image I with a natural language expression S are given as input. $\hat{M}$ is the pixel-wise prediction mask, and M is the corresponding real pixel-wise segmentation mask. Here $\hat{M} \in (0, 1)$ represents the foreground probability of a pixel and $M \in \{0, 1\}$, where 1 means the pixel is referred to by S while 0 means the background. The goal of this task is to segment the corresponding region in image I referred by the expression S. To achieve the goal, the key problem is how to understand both visual and textual input and successfully separate the target region from others. In this paper, we first propose a semantic guided approach to enhance the image referring segmentation by adding multi-scale features and semantic embedding. To refine the prediction produced by the segmentation network, we further model the network into a generative adversarial framework by introducing a novel adversarial loss term. In this framework, we regard the segmentation network as a generator which generates the masks and the added discriminator is employed to distinguish those masks from the ground truth ones. Formally, our proposed ASGN is to train a network $f(I, S; \theta)$ parameterized by θ. The network parameter θ is optimized by a pixel-wise classification loss and an adversarial loss:

$$\min_{\theta}(\mathcal{L}_{\text{seg}}(f(I, S; \theta)) + \mathcal{L}_{\text{adv}}(f(I, S; \theta))), \tag{1}$$

where $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{adv}}$ are the loss functions of pixel-wise segmentation and adversarial loss terms, respectively.

An overview of the proposed ASGN architecture is shown in Fig. 2. Our proposed framework mainly consists of two parts: the generator (left) and the discriminator (right). Specifically, the generator contains two components, i.e. semantic guidance segmentation mask prediction and LSTM-based natural language expression encoder.

B. Segmentation Mask Prediction

First, we describe the generator which is employed to produce the predicted masks. Following previous work [2], our basic architecture includes the visual feature extractor and a sentence encoder which are used to produce the image and sentence representation, respectively. The visual extractor is built upon the ResNet-101 or DeepLab-101. To allow the model to reason about spatial relationships such as "person on the right", 8 spatial coordinates are also concatenated with the extracted features. Then the concatenated tensor is sent to two additional convolutional layers to predict the final segmentation mask. Given the ground truth binary segmentation mask, the pixel-wise classification loss function is defined as:

$$\mathcal{L}_{CE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} L\left(M^{ij}, \hat{M}^{ij}\right), \tag{2}$$

where i and j are the spatial coordinates, W and H are the width and height of the predicted mask, with $1/8$ size of the original input image. This is set according to the downsampling rate (8) of the segmentation network. Specifically, the per-pixel cross entropy loss $\mathcal{L}_{CE}$ in Equation 2 can be written as follows:

$$\begin{aligned}\mathcal{L}_{CE} = -\frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \Big(& M^{ij} \log(\hat{M}^{ij}) \\ & + (1 - M^{ij}) \log\left(1 - \hat{M}^{ij}\right)\Big)\end{aligned} \tag{3}$$

Even though the pixel-wise prediction generated from the basic network performs well in previous work, it is still ambiguous in marginal regions that cannot align with the actual border. For instance, the body of a person is easy to segment but the border of the arms and legs is hard to align with the actual borders completely. To address the above issues, we propose to add more

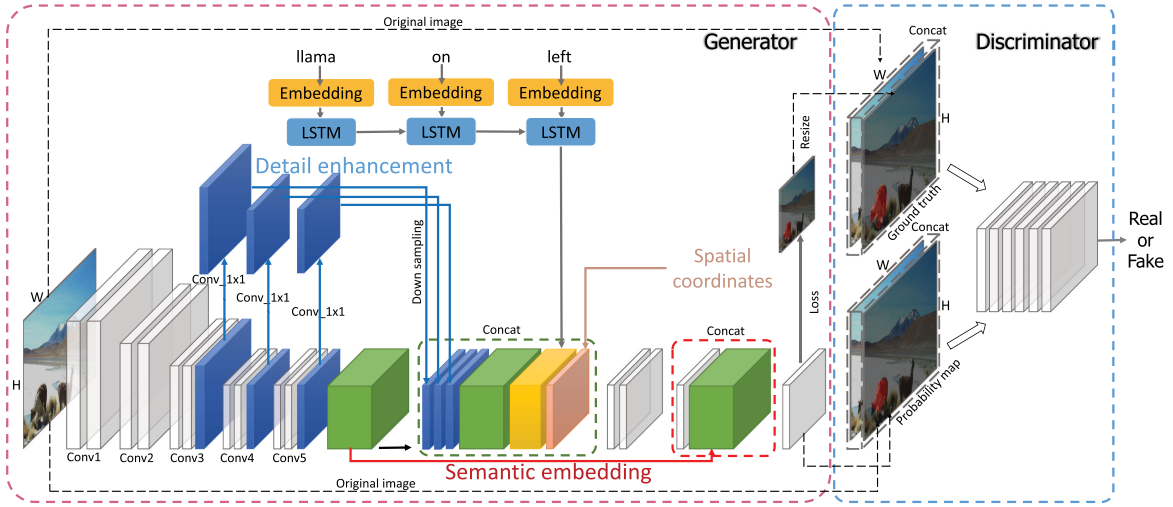


Fig. 2. Network architecture of the proposed ASGN model. It mainly consists of two parts: the generator (left) and the discriminator (right).The generator employs the fully convolutional networks as the backbone to extract image features. We extract feature maps from the network backbone (conv3, conv4 and conv5) as the multi-scale features (blue connection) to enhance the details and then concatenate them with image feature (green block), language feature from LSTM (yellow block) and spatial coordinates (orange block). Then, the fused features are fed into the following two convolutional layers. The semantic features (red connection) are concatenated to the final features before the mask prediction to reduce the ambiguity introduced by word embedding. After generating predictions, a discriminator is introduced to relieve the distribution inconsistencies. In the discriminator network, we first re-size the predictions to the same size as the original images and then concatenate the original image with the predicted masks and real masks respectively. After that, these two image stacks are fed into a classification network with six layers.

detailed semantic guidance to improve the edges of the target region. Particularly, we apply two types of skip layer connections as shown in Fig. 2: Detail enhancement (the blue connection in Fig. 2) and Semantic embedding (the red connection in Fig. 2).

1) Detail Enhancement: In general, CNNs usually employ several hidden layers to hierarchically learn multi-level representation of images. In this paper, we propose to extract the multi-scale features and concatenate them to the feature fusion block (highlighted using circled green dotted line in Fig. 2), to enhance the image details. This is achieved by a skip connection (blue connection in Fig. 2) from the backbone to the fusion block with the intermediate convolutional layers. Specifically, the features from the conv3, conv4 and conv5 layer of the backbone are extracted as the multi-scale feature guidance (more details on the selection of these layers please refer to the experiments section). Additionally, we apply three 1×1 convolution layers on the extracted feature maps to reduce the number of channels for purpose of avoiding outweighing the importance of the rich features. Then, the features are bilinearly downsampled to the same size of those to $1/8$ size of the original input image, which are later concatenated for segmentation.

2) Semantic Embedding: In existing works, the inferring segment mask is predicted directly from the fusion block that combines image and natural language features. However, the natural language feature may introduce semantic ambiguities, e.g., foreground and background with similar expression descriptions. To better model the semantic information, we propose to use a semantic embedding to provide further guidance for the referring mask prediction, as shown in Fig. 2 (highlighted using circled red dotted line). It concatenates the high-level feature map produced by FCN with the probability map and then the final masks will be predicted after a convolutional layer.

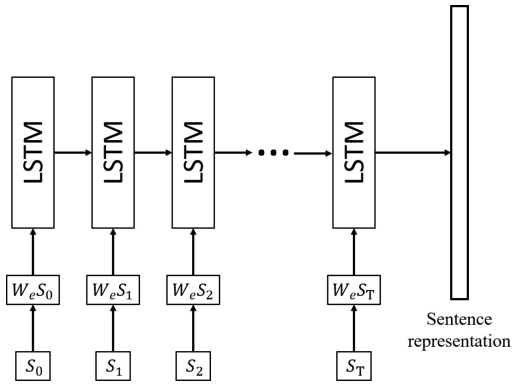


Fig. 3. The structure of the LSTM-based language encoder. This model takes the embedded word after a word embedding matrix as input at each time step.

By incorporating the semantic embedding, the predicted referring mask is improved both quantitatively and qualitatively (refer to the experiment for details).

C. LSTM-Based Language Expression Encoder

To represent the input natural language expression, we adopt a recurrent Long-Short Term Memory (LSTM) network to encode the text sequences to fixed-length vectors [48], [49]. Specifically, we first fix each text sequence to a uniform length by padding with zero or cut the excess. The structure of the encoding network is illustrated in Fig. 3, we denote the trimmed sentence by $S = (S_0, \ldots, S_N)$ where we represent each word as a one-hot vector S_t (the t-th word of the vector). After using a word embedding matrix W_e, we put the embedded word vector $W_e S_t$ into the LSTM network at each time step. The LSTM network

includes various gate mechanisms including input gate i_t, forget gate f_t, output gate o_t, memory state c_t and hidden state h_t. The definition of the gates and states are as follows:

$$\begin{aligned} i_t &= \sigma\left(W_{ix}x_t + W_{ih}h_{t-1}\right) \\ f_t &= \sigma\left(W_{fx}x_t + W_{fh}h_{t-1}\right) \\ o_t &= \sigma\left(W_{ox}x_t + W_{oh}h_{t-1}\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1}) \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{4}$$

where $\odot$ represents the element-wise multiplication and the W matrices are the learned parameters. The hidden state h_T at the last time step aggregates the whole sentence and is regarded as the representation of the language expression.

D. Generative Adversarial Network Based Framework

In this section, we elaborate the generative adversarial module. The cross-entropy loss of pixel-wise classification calculates the difference between the network prediction and the ground truth labels independently. It tends to ignore the label statistics. For instance, the shape of a target region or object cannot be captured by the pixel-wise loss function. As a result, in addition to the cross-entropy loss $\mathcal{L}_{CE}$, we introduce another adversarial loss term that is based on the GAN [17]. Since the adversarial training captures the distribution of the entire image, mismatches in the label statistics can be penalized by the adversarial loss term. In traditional GAN, a generative model G and a discriminative model D are jointly trained to play a minimax game. G maps samples from noise distribution P_z to the data distribution P_{data}, while D aims to distinguish the prediction from G and the ground truth. Consequently, the G tries to predict data that as "real" as possible.

The objective function of this game is:

$$\min_G \max_D \mathbb{E}_{x\sim P_{data}}\left[\log D(x)\right] + \mathbb{E}_{z\sim P_z}\left[\log(1 - D(G(x))\right] \tag{5}$$

Different from the traditional GAN that samples from noise, our adversarial loss term is conditioned on the predicted segmentation mask, as shown in Fig. 2. Specifically, we adapt the G to our mask prediction function F and condition on the concatenation of both the mask $\hat{M} = F(x)$ and input x. We empirically found that better results can be achieved compared to traditional GAN. Therefore the proposed adversarial loss term $\mathcal{L}_{adv}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = &\max_D \mathbb{E}_{x\sim P_{data}}[\log D(x \oplus M)] \\ &+ \mathbb{E}_{x\sim P_{data}}[\log(1 - D(x \oplus F(x))], \end{aligned} \tag{6}$$

where $\mathbb{E}$ is the empirical estimate of the expected value of the probability and x is the input image. The operator $\oplus$ represents concatenation. We adapt this framework for referring image segmentation by jointly optimizing F and D.

Then the final joint loss function can be defined as a weighted aggregation of the pixel-wise loss term and the adversarial loss term. λ is referred to as the balancing weight, and the joint loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{CE}, \tag{7}$$

where $\mathcal{L}_{CE}$ and $\mathcal{L}_{\text{adv}}$ refer to $\mathcal{L}_{\text{seg}}(f(I, S; \theta))$ and $\mathcal{L}_{\text{adv}}(f(I, S; \theta))$ in (1).

During the inference, we only use the generator network for the referring image segmentation inference. As a result, we can refine the prediction without any additional computation and time after the training phase. Given a test image, the Softmax layer of the generator outputs the probabilities of each pixel belonging to the referring object. Then the pixel with probability larger than threshold δ is assigned to the referring object.

IV. Experiments

In this section, we evaluate the proposed ASGN model on four benchmark datasets by comparing with the previous referring image segmentation methods. In addition, further evaluation is performed by applying our model to different baseline methods.

A. Datasets

Both UNC [7] and UNC+ [7] are constructed based on MS COCO dataset [57]. The difference between these two datasets is that no location words are allowed in the expression in the UNC+ dataset. UNC includes 142,209 referring expressions and 19,994 images, while UNC+ includes 141,564 expressions and 19,992 images. We use the same data split as in [7]. For UNC dataset, we use 120,624 image/expression pairs as the training set and employ tree testing subsets with 10,834, 5,657 and 5,095 pairs each. Similarly, for the UNC+ dataset, the size of training and testing sets are 120,191 and (10,758, 5,726, 4,889), respectively.

Containing 104,560 expressions and 26,711 images, Google-Ref [48] is also selected from MS COCO dataset [57]. We use the same data split as in [48]. 85,474 image/expression pairs are used for training and 9,536 pairs for testing.

ReferItGame [58] contains 130,525 expressions and 19,894 natural images. Different from the above three datasets, ReferItGame contains background segmentation masks, such as "sky" and "water". We use the same data split as in [1]: 59,976 image/expression pairs for training and 60,105 pairs for testing.

B. Evaluation Metrics

As the region-based metric Intersection over Union (IoU) [1], [2] takes into account both the false and the missed values for each class, it has been used as a standard metric for semantic segmentation evaluation. Here we also utilize the IoU as the evaluation metric in our experiments.

To allow for comparison, we also evaluate with the Precision@X metric at 5 different IoU thresholds from easy to hard: 0.5, 0.6, 0.7, 0.8, 0.9 which is consistent with previous work [1], [2]. The Precision@X metric is the percentage of test samples if the IoU between prediction and ground-truth passes the threshold. e.g. Precision@0.5 is the percentage of testing samples if its predicted segmentation overlaps with the ground-truth region by at least 50% IoU.

TABLE I
COMPARISON WITH BASELINE MODELS ON IOU. IN THE FIRST COLUMN, R MEANS RESNET WEIGHTS, D MEANS DEEPLAB WEIGHTS, DE MEANS DETAIL ENHANCEMENT, SE MEANS SEMANTIC EMBEDDING, AT MEANS ADVERSARIAL TRAINING, DCRF MEANS DENSECRF. BLANK ENTRIES WHERE AUTHORS DO NOT REPORT PERFORMANCE

Method	UNC			UNC+			Google-Ref	ReferItGame
	val	testA	testB	val	testA	testB	val	test
Hu *et al.* [53] (2016)	-	-	-	-	-	-	34.06	49.91
D+RMI+DCRF [2] (ICCV 2017)	45.18	45.69	45.57	29.86	30.48	29.50	34.52	58.73
LBIE [54] (CVPR 2018)	-	-	-	-	-	-	-	50.09
DMN [55] (ECCV 2018)	49.78	54.83	45.13	38.88	44.22	32.29	36.76	52.81
KWAN [56] (ECCV 2018)	-	-	-	-	-	-	36.92	59.09
SNLE[1]	-	-	-	-	-	-	-	48.03
SNLE+AT	-	-	-	-	-	-	-	50.17
SNLE+DE+SE+AT	-	-	-	-	-	-	-	**52.87**
R+LSTM	39.29	39.80	39.15	26.87	27.52	25.28	28.06	54.04
R+LSTM+AT	41.18	42.19	40.30	30.14	31.44	29.37	34.42	54.41
R+LSTM+DE+SE+AT	**45.37**	**46.84**	**45.15**	**34.38**	**35.76**	**32.49**	**36.82**	**57.92**
D+LSTM	43.32	43.71	43.25	28.61	29.03	28.14	32.86	56.78
D+LSTM+AT	44.74	45.75	44.05	33.08	33.59	31.27	38.89	57.11
D+LSTM+DE+SE+AT	**48.04**	**48.86**	**47.64**	**36.25**	**37.47**	**34.48**	**40.36**	**58.43**
D+RMI	44.51	44.83	44.74	30.18	31.02	30.05	34.27	57.39
D+RMI+AT	45.40	46.10	44.87	33.88	35.01	32.70	39.60	58.74
D+RMI+DE+SE+AT	**50.46**	**51.20**	**49.27**	**38.41**	**39.79**	**35.97**	**41.36**	**60.31**

C. Implementation Details

Several baseline methods are compared with the proposed ASGN framework. R/D+LSTM are the baseline models which employ ResNet-101 or DeepLab-101 to extract image features respectively. D+RMI [2] uses a recurrent multimodal LSTM (mLSTM) interaction model to extract text features instead of LSTM. SNLE represents the method proposed by [1] which employs the same architecture as the baseline model R+LSTM, except that it uses FCN-32s to extract image features. DE, SE and AT represent the two skip connections of detail enhancement, semantic embedding and the adversarial training respectively. Specifically, configurations without the AT indicate that only the $\mathcal{L}_{CE}$ term in Equation 7 is used for training, while the methods with AT means both the two terms of $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{CE}$ are used. However, only using the $\mathcal{L}_{\text{adv}}$ term will lead to a large performance decrease, since this loss term plays a supporting role in the segmentation network training. So only with and without second term experiment results are provided.

Following [2], in our implementation, the size of the input image and ground truth segmentation are set to 320×320. The dimensions of image feature and sentence vector are 1,000. δ is set to 10^{-9}. The backbone ResNet-101 is pretrained on ImageNet [59], and DeepLab-101 [25] is finetuned on Pascal VOC [33]. To optimize our networks, we follow the standard approach from [17]: the model is trained alternatively by one gradient descent step on D, then one step on G. We use the Adam optimizer with a fixed learning rate of 0.00025 on both generative phase and discriminative phase.

D. Comparison to State-of-the-Art

We compare our proposed ASGN approach to baseline methods and state-of-the-art methods on four datasets. The main results of our evaluation are summarized in Table I. The state-of-the-art methods are listed in chronological order. It can be observed that our proposed method outperforms all the baseline methods by a large margin in terms of the overall IoU metric, as shown in the first part of Table I.

We show the impact of GAN and semantic guidance separately, in the following parts of Table I. The second row of each baseline shows the performance of adversarial training (+AT) and the third row indicates the superiority of feature embedding (+DE+SE). In particular, R+LSTM+AT and D+LSTM+AT achieve 34.42% and 38.89% in Google-Ref dataset, which outperforms the baseline models more than 6%. Also, our AT models on UNC+ dataset achieve 4–5% performance improvement on the validation and test sets. In comparison to the performance on ReferItGame dataset, the improvement by using the adversarial model is relatively lower than the other datasets. One possible reason for the smaller performance gain is: the ReferItGame dataset contains more "stuff" segmentation masks, i.e. "sky" and "ground", where these "stuff" masks are different from other objects since "stuff" has various shapes and different pixel label statistics. It is much more difficult for the discriminator to distinguish the predictions from real masks.

Compared with the +AT setups, we can see that our final model in the third row outperforms prior methods by a relatively large margin. In particular, R+LSTM+DE+SE+AT leads to 45.37%, 46.84% and 45.15% for three testing sets of UNC database, with significant improvements (more than 4%) over the second row. The results show that after visual feature embedding, the model can depict the target region more accurately.

In addition, we show the corresponding qualitative results in Fig. 4 as well. The results in Fig. 4 show that by adding detail semantic guidance and adversarial training scheme, ASGN model can segment the region described by expressions well. It is clear that the segmentation results in the fourth column cover better target objects. Besides, in the last column, after adding multi-scale features and semantic embedding, more semantic information is introduced into the model which can well depict

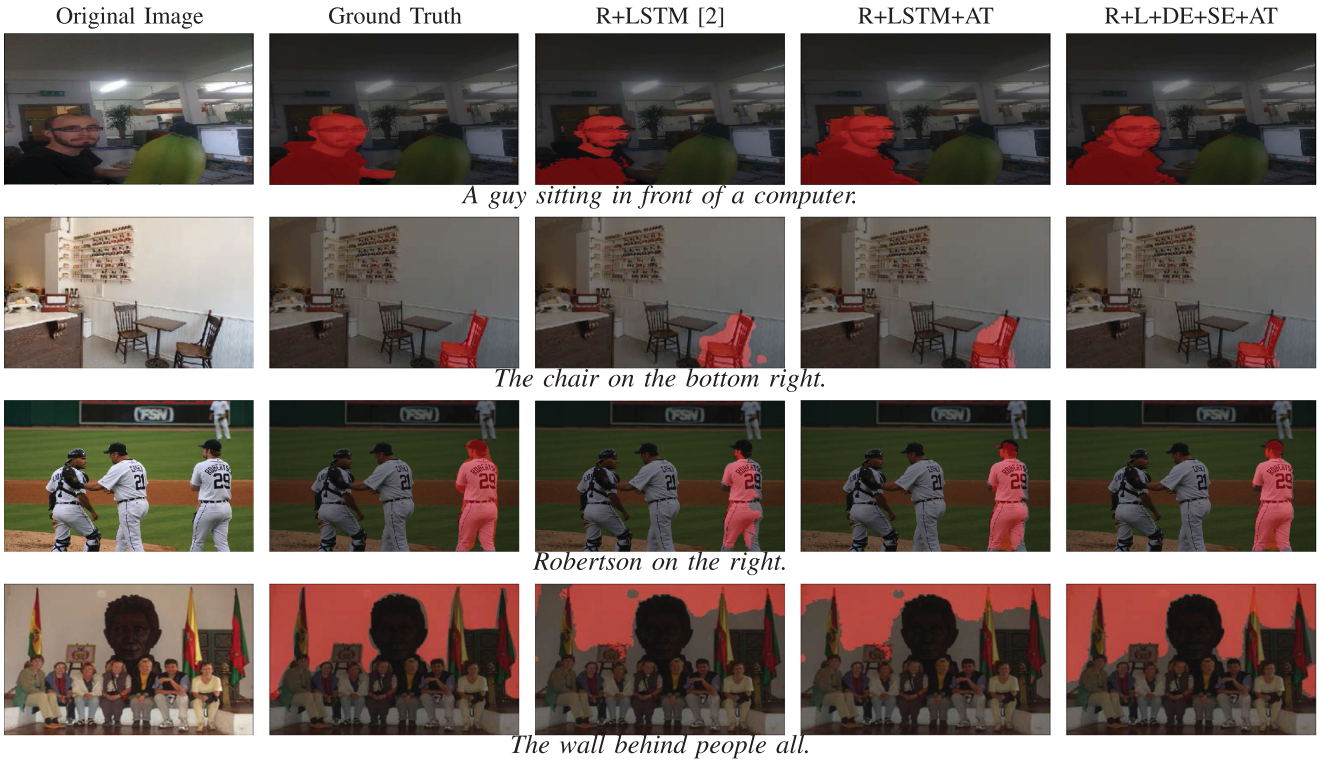


Fig. 4. Visualization results of referring image segmentation. We compare our two proposed models with baseline R+LSTM [2]. The fourth and fifth column are our proposed approaches.

Fig. 5. Visualization of ablation study on UNC dataset. The last four columns are our proposed approaches.

the object boundaries. The ReferItGame dataset is more flexible compared to the other datasets as it contains some background region like "sky" in addition to the foreground objects. An example is shown in the last row of Fig. 4, in which the proposed ASGN model can well segment the "wall" in the images. The experimental results demonstrate that our model works well for both foreground and background regions.

Comparing to the state-of-the-art methods, the results of the proposed approach outperform existing works. As demonstrated above, the performance can be significantly improved by our approach. Although the performance of the proposed framework on four example baselines cannot exceed the state-of-art methods on all datasets (i.e. testA of UNC), our approach is general and can be embedded in any other state-of-the-art frameworks to achieve a further improvement, since most of referring image segmentation methods have similar segmentation network backbone (ResNet or DeepLab) to our given baseline examples.

E. Ablation Study

To validate the effectiveness of each component in our proposed ASGN approach, we conduct ablation studies on UNC dataset with ResNet+LSTM baseline method. Quantitative and qualitative results are shown in Table II and Fig. 5, respectively.

Semantic Embedding. The introduced semantic feature embedding encourages the predicted mask to conclude more semantic information which can counteract the impact caused by adding natural language features. To analyze this component, we conduct experiments on with and without semantic feature embedding on the UNC dataset. As the results shown in

TABLE II
RESULTS OF EACH METRIC TERM OF RESNET MODELS ON UNC DATASET. IN THE FIRST COLUMN, R MEANS RESNET WEIGHTS, DE MEANS DETAIL ENHANCEMENT, SE MEANS SEMANTIC EMBEDDING AND AT MEANS MODEL WITH ADVERSARIAL TRAINING

Testing Set	Method	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	IoU
val	R+LSTM	31.99	19.37	8.72	2.23	0.08	39.29
	R+LSTM+AT	34.75	20.35	8.45	2.10	0.09	41.18
	R+LSTM+SE	34.88	22.11	10.55	3.21	0.11	40.14
	R+LSTM+SE+AT	35.76	22.19	10.68	3.30	0.16	42.42
	R+LSTM+DE	41.51	30.44	19.06	7.96	0.94	44.88
	R+LSTM+DE+AT	42.25	31.56	20.35	8.94	1.01	45.06
	R+LSTM+DE+SE	42.13	30.03	19.40	8.20	0.82	44.92
	R+LSTM+DE+SE+AT	43.88	33.76	22.30	10.95	1.05	45.37
testA	R+LSTM	31.13	19.00	8.66	2.14	0.05	39.80
	R+LSTM+AT	34.52	21.00	9.23	2.02	0.04	42.19
	R+LSTM+SE	34.82	21.95	9.79	3.02	0.05	40.69
	R+LSTM+SE+AT	37.79	22.56	10.09	2.44	0.11	43.70
	R+LSTM+DE	42.46	31.25	20.40	8.38	0.64	45.53
	R+LSTM+DE+AT	43.74	32.13	22.12	11.98	1.33	46.17
	R+LSTM+DE+SE	43.91	30.88	21.38	9.93	0.82	45.63
	R+LSTM+DE+SE+AT	44.05	34.81	24.71	13.77	2.33	46.84
testB	R+LSTM	33.58	20.96	9.91	2.37	0.16	39.15
	R+LSTM+AT	34.39	22.04	10.30	2.47	0.10	40.30
	R+LSTM+SE	35.54	23.00	11.25	3.49	0.21	39.86
	R+LSTM+SE+AT	37.19	23.96	11.23	2.61	0.18	42.13
	R+LSTM+DE	42.32	32.33	20.71	9.54	1.12	44.71
	R+LSTM+DE+AT	43.37	32.38	20.99	10.53	1.33	44.97
	R+LSTM+DE+SE	42.16	32.97	20.86	9.38	1.13	44.74
	R+LSTM+DE+SE+AT	44.08	33.25	21.67	11.54	2.08	45.15

Table II, with semantic feature works better than without such component. In particular, R+LSTM+SE achieves 40.14% in verification set. All those experiments results demonstrate the effectiveness of semantic feature embedding for referring segmentation task. In Fig. 5, the fifth column (R+LSTM+SE) of each row shows the corresponding qualitative performance. We can observe that without semantic embedding the model fails to well segment the object boundaries while with semantic embedding the model can predict better segmentation mask for the inferred objects. For example, in the first row in Fig. 5, the edges of the man's shoulders, arms and head can be detected well by the R+LSTM+SE approach compared to the third column (R+LSTM) method.

Detail Enhancement. Based on the baseline model, we add the multi-scale feature extracted from the front convolutional layers (conv3, conv4 and conv5) to facilitate the segmentation. To find out how to structure the detail enhancement skip connection, we investigate the effect by varying convolutional layers combination for providing detail information. As shown in Table III, all the performances are improved compared to the baseline R+LSTM method, which validates the effectiveness of the detail enhancement. We observe that the IoU drops by near 3% when only using conv1 layer compared to other combinations. The reason is that the first layer (conv1) captures too detailed local information which does not contain semantic information like object contours. On the other side, all the combinations show better performance compared to single layer settings, due to the abundant information captured. More than 5% performance improvement can be brought when employing the conv3, conv4 and conv5. In consequence, we perform the detail enhancement from the combination of conv3, conv4 and conv5 features for better segmentation performance.

TABLE III
COMPARISON OF DIFFERENT CONVOLUTIONAL LAYER OR LAYER COMBINATION IN DETAIL ENHANCEMENT ON UNC DATASET

Methods	val	testA	testB
R+LSTM	39.29	39.80	39.15
conv1	41.56	42.28	41.23
conv2	42.33	43.18	41.70
conv3	42.21	43.23	41.70
conv1+conv2+conv3	42.76	42.97	42.41
conv4+conv5	43.56	43.51	43.43
conv3+conv4+conv5	**44.88**	45.53	**44.71**
conv2+conv3+conv4+conv5	44.55	**45.65**	43.77
conv1+conv2+conv3+conv4+conv5	43.96	44.87	43.40

We now proceed to evaluate the detail enhancement and investigate how it benefits the referring segmentation. For detail enhancement, we can see that both the overall IoU and Precision@X metrics in Table II of the ASGN method are higher than the baseline approach. In particular, R+LSTM+DE improves the IoU from 39.29% to 44.88%, which outperforms the baseline model by more than 5%. Therefore, the multi-scale feature maps from front layers of network backbone can introduce more visual detailed information that is beneficial to referring image segmentation. The corresponding qualitative performance of this component is shown in the sixth column (R+LSTM+DE) in Fig. 5. It can be observed that the model with the multi-scale feature connection can segment the corresponding region with explicit edges. In particular, in the last row in Fig. 5, after introducing the detail enhancement skip connection, our R+LSTM+DE approach can well predict the boundaries of the bear, especially the edges of its arms and legs. This means the multi-scale features enrich the detailed visual information by the skip connection.

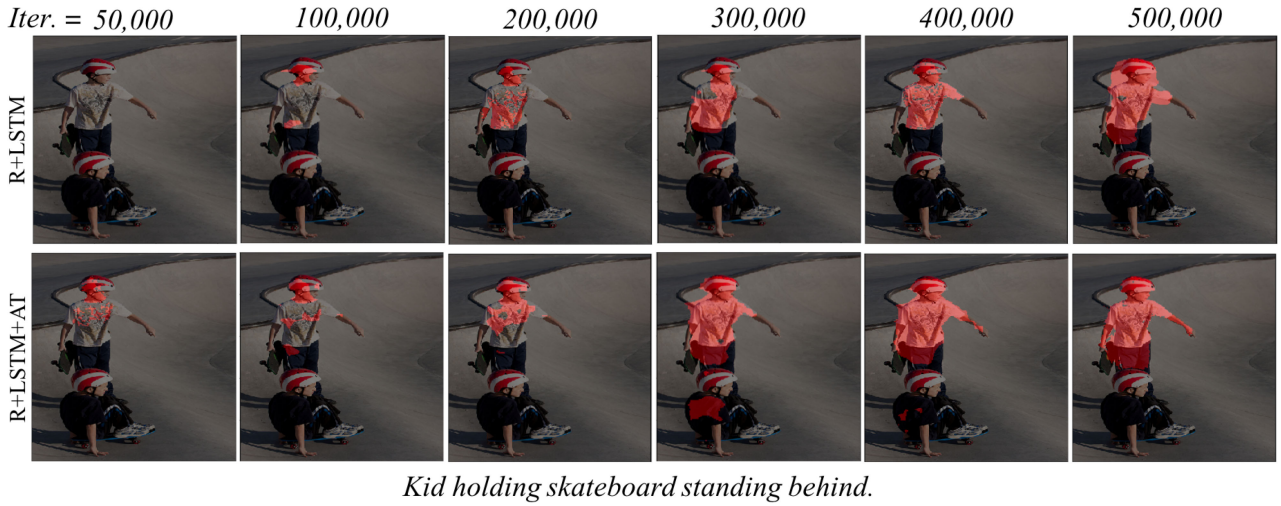


Fig. 6. Visualization of the improvement about prediction with the increasing of iterations. In each example, the first row is the result of R+LSTM, the second row is the result of R+LSTM+AT.

TABLE IV
COMPARISON OF THE CLASSIFICATION ACCURACY WITH OR WITHOUT GAN. WE TRAIN TWO SVM CLASSIFIERS TO DIVIDE MASKS INTO THE PREDICTED CLASS GENERATED BY TWO METHODS AND THE GROUND TRUTH RESPECTIVELY

Methods	Accuracy
R+LSTM	92
R+LSTM+AT	90.75

Adversarial Training. We also analyze the effectiveness of the adversarial training in our proposed referring image segmentation method. As shown in Table IV, we select 2,000 images from the Google-Ref dataset and use two methods of R+LSTM and R+LSTM+AT to predict the corresponding mask inferences. After that, we mix the ground truth masks with the predicted masks together. Then we extract visual features of the two mixed sets from the fc7 layer of AlexNet [60]. Then two SVM classifiers are trained (half size as the training set) to separate the predicted masks from the ground truth. The binary classification accuracy of R+LSTM features is 92%, while the accuracy for R+LSTM+AT is 90.75%. The lower accuracy when adding the GAN indicates that the proposed adversarial loss leads the prediction to be closer to the ground truth. That is, our network successfully predicts mask that is less distinguishable from the ground truth.

The quantitative results are shown in Table I and Table II. As mentioned above, the performance of methods with +AT setups shows superiority under overall metrics. In Table II, we present the results of with AT on different compared methods including embedding schemes. In particular, on the testA sets of UNC database, the IoU is increased from 39.80% to 42.19% for R+LSTM. Similar performance gain can be observed in adding the embedding schemes, with more than 3% performance improvement (from 40.69% to 43.70%). We also show the corresponding qualitative performance of the adversarial component in the fourth column (R+LSTM+AT) in Fig. 5. It can be observed that the segmentation maps produced by R+LSTM+AT method are more similar to the ground truth than the ones predicted by R+LSTM method.

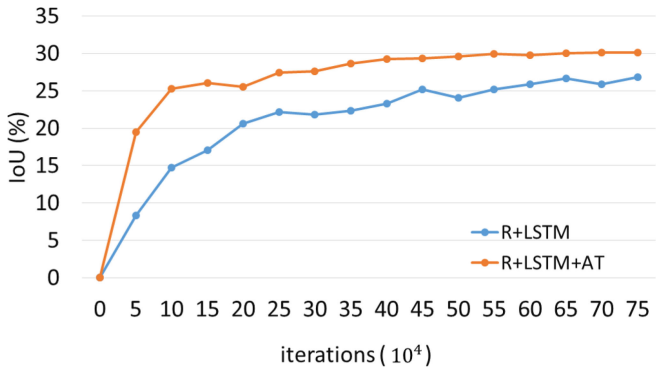


Fig. 7. IoU across training iterations on the UNC dataset on validation data with and without adversarial training.

In Fig. 7, we display the evolution of the referring segment prediction accuracy on the UNC dataset, using the baseline model and the corresponding adversarial model. Note that the adversarial strategy results in less over-fitting, i.e., generating a regularization effect, leading to high accuracy on validation data. It is worth mentioning that our ASGN framework can obtain a better segment result with fewer iterations and converges faster compared to previous methods. In Fig. 6, we further visualize the improvement of prediction with the increasing of iterations. We can see that the model with adversarial training results in better performance compared to the one without GAN, at the same iterations.

We also evaluate the computational time and memory costs of the baseline model and the adversarial training model. All the computational time experiments are performed on a PC with GeForce GTX 1080 GPU. The results of testA in UNC dataset are shown in Table V. It can be observed that the proposed framework can achieve comparable testing time and memory

TABLE V
COMPARISON OF COMPUTATIONAL TIME AND MEMORY COSTS

Methods	IoU	training time(h)	testing time(s)	memory costs(m)
R+LSTM	39.80	19.46	652.25	448.46
R+LSTM+DCRF	40.44	19.46	2782.04	448.46
R+LSTM+DE+SE+AT (Proposed)	46.84	39.21	660.04	481.23

TABLE VI
COMPARISON OF DIFFERENT λ SETTINGS ON UNC DATASET

λ	val	testA	testB
0.01	39.50	39.82	40.44
0.1	**45.37**	**46.84**	**45.15**
1	42.35	43.37	41.97

(a) The light on the left in the air above the head of woman.

(b) The window in the left.

Fig. 8. Failure cases of our model. Given the input images (left) and corresponding referring expressions (below), the predicted referred regions by our model (right) and the ground truth segment masks (middle) are shown.

cost to the baseline (R+LSTM). Compared with methods using post-processing like DCRF, the ASGN approach can save more time in test phase, which is more suitable for practical applications. Although the proposed network increases the training time cost, it is valuable and acceptable because of the nearly 7% improvement in IoU.

In summary, the extensive experimental results indicate that the proposed ASGN model is able to boost the performance of referring semantic segmentation by a large margin.

Parameter Setting. The value of λ in Equation (7) and threshold δ are empirically set according to the experiment on UNC dataset with ResNet+LSTM baseline method. We select three λ values of $0.01, 0.1$ and 1. Since the pixel-wise cross entropy loss is much larger than the adversarial loss, we make $\lambda \leq 1$ to balance the training. As shown in Table VI, compared with $\lambda = 0.01$ and 1, the precision outperforms by a relatively large margin when $\lambda = 0.1$. Therefore we choose the λ to be 0.1 according to the experiment. Then we test the value of parameter δ and select six values from 10^{-10} to 10^{-5}. The results of different values are comparable. Thus we choose $\delta = 10^{-9}$ in all the evaluations following previous work [1], [2].

Failure Case. Typical failure cases are depicted in Fig. 8. We find that the failure cases are mainly about difficult small objects and ambiguous referring. In Fig. 8(a) the network cannot segment the referring "light", which is too small to segment. It may be alleviated by enlarging the scale of input images. In Fig. 8(b), the network segments the wrong window since the correct corresponding region is the window on another building outside the large window in the current building. It is mainly caused by the ambiguity in the expression which confuses the model.

V. CONCLUSION

In this paper, we proposed an adversarial semantic guidance network for referring image segmentation. It not only encourages the distributions of the network inference and the ground truth to be similar but also adds more detailed semantic guidance. We first leverage multi-scale features by a skip connection so that more detailed visual information is introduced. Then semantic embedding is utilized to eliminate the impact from the combination of expression features. Finally, in order to constrain the distribution similarity, we further introduce the adversarial training scheme as a supplemental loss term in addition to the cross-entropy loss. Extensive experimental results demonstrate that the proposed ASGN approach leads to improvements in referring image segmentation on various benchmark datasets. In addition, the proposed framework is shown to be beneficial to other existing referring image segmentation models.

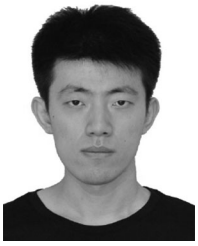
REFERENCES

[1] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 108–124.
[2] C. Liu *et al.*, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1271–1280.
[3] G. P. Laput *et al.*, "Pixeltone: A multimodal interface for image editing," in *Proc. Special Interest Group Comput.-Human Interact. Conf. Human Factors Comput. Syst.*, 2013, pp. 2185–2194.
[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.
[5] 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. 2018.
[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[7] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 69–85.
[8] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 792–807.
[9] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4995–5004.
[10] X. Li and S. Jiang, "Bundled object context for referring expressions," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2749–2760, Oct. 2018.
[11] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[12] J. Dong, X. Li, and C. G. M. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, Dec. 2018.
[13] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
[14] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.
[15] T. Wang, Z. Ji, Q. Sun, Q. Chen, and X. Y. Jing, "Interactive multilabel image segmentation via robust multilayer graph constraints," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2358–2371, Dec. 2016.
[16] J. Tighe, M. Niethammer, and S. Lazebnik, "Scene parsing with object instances and occlusion ordering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 3748–3755.
[17] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[18] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst. Workshop Adversarial Training*, Barcelona, Spain, 2016.
[19] Y. Wei *et al.*, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1568–1576.
[20] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2746–2754.
[21] H. Xiao, Y. Wei, Y. Liu, M. Zhang, and J. Feng, "Transferable semi-supervised semantic segmentation," in *Proc. AAAI*, pp. 7420–7427, 2018.
[22] J. Jiao *et al.*, "Geometry-aware distillation for indoor semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2869–2878.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015. pp. 1–14.
[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[25] 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. 2018.
[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, Apr. 2016.
[27] Y. Wei *et al.*, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7268–7277.
[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 801–818.
[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2881–2890.
[30] J. Li, P. Yuan, D. Gu, and Y. Tian, "Hierarchical deep co-segmentation of primary objects in aerial videos," *IEEE MultiMedia*, vol. 26, no. 3, pp. 9–18, Jul.–Sep. 2019.
[31] L. Zhang, C. Fu, and J. Li, "Collaborative annotation of semantic objects in images with multi-granularity supervisions," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 474–482.
[32] B. Kang, Y. Lee, and T. Q. Nguyen, "Depth adaptive deep neural network for semantic segmentation," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2478–2490, Sep. 2018.
[33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, pp. 303–338, 2010.
[34] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *Comput. Sci.*, vol. 3, pp. 469–477, 2015.
[35] W. Wang and J. Shen, "Higher-order image co-segmentation," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1011–1021, Jun. 2016.
[36] Z. Lou and T. Gevers, "Extracting primary objects by video co-segmentation," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2110–2117, Dec. 2014.
[37] C. Wang, Y. Guo, J. Zhu, L. Wang, and W. Wang, "Video object co-segmentation via subspace clustering and quadratic pseudo-boolean optimization in an MRF framework," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 903–916, Jun. 2014.
[38] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[39] S. Reed *et al.*, "Generative adversarial text to image synthesis," in *Proc. IEEE Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.
[40] H. Zhang *et al.*, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5908–5916.
[41] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2536–2544.
[42] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 597–613.
[43] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 517–532.
[44] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.
[45] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Advances Neural Inform. Process. Syst.*, 2017, pp. 700–708.
[46] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[47] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 702–716.
[48] J. Mao *et al.*, "Generation and comprehension of unambiguous object descriptions," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 11–20.
[49] R. Hu *et al.*, "Natural language object retrieval," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4555–4564.
[50] J. Mao *et al.*, "Deep captioning with multimodal recurrent neural networks (M-RNN)," in *Proc. Int. Conf. Learn. Represent.*, pp. 1000–1020, 2015.
[51] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2625–2634.
[52] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele, "Grounding of textual phrases in images by reconstruction," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 817–834.
[53] R. Hu, M. Rohrbach, S. Venugopalan, and T. Darrell, "Utilizing large scale vision and text datasets for image segmentation from referring expressions," 2016, *arXiv:1608.08305*.
[54] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8721–8729.
[55] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 630–645.
[56] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 38–54.
[57] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.
[58] 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.*, 2014, pp. 787–798.
[59] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.
[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

Shuang Qiu is currently working toward the Ph.D. degree with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. Her research interests include cross-media retrieval, computer vision, and deep learning.

Yao Zhao received the B.S. degree from Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor with BJTU in 1998 and a Full Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is currently leading several national research projects from the 973 Program, 863 Program, and National Science Foundation of China. He serves on the editorial boards of several international journals, including as an Area Editor for *Signal Processing: Image Communication* (Elsevier), and as an Associate Editor for *Circuits, System & Signal Processing* (Springer). He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010.



Jianbo Jiao received the Ph.D. degree from City University of Hong Kong, Hong Kong. He is currently a Postdoctoral Researcher with the Department of Engineering Science, University of Oxford, Oxford, U.K. He was a Visiting Scholar with Beckman Institute, University of Illinois at Urbana-Champaign. His research interests include computer vision and machine learning.

Yunchao Wei received the Ph.D. degree from Beijing Jiaotong University, Beijing, China, in 2016, advised by Prof. Yao Zhao. He is currently a Postdoctoral Researcher with Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, IL, USA, working with Prof. Thomas Huang. His current research interest focuses on computer vision techniques for large-scale data analysis. Specifically, he has done work in weakly- and semi-supervised object recognition, multi-label image classification, video object detection, and multi-modal analysis. He received Excellent Doctoral Dissertation Awards of Chinese Institute of Electronics (CIE) and Beijing Jiaotong University in 2016, the Winner prize of the object detection task (1a) in ILSVRC 2014, the Runner-up prizes of all the video object detection tasks in ILSVRC 2017, the Winner Prizes of all human parsing tracks in the 2nd LIP challenge.

Shikui Wei received the Ph.D. degree in signal and information processing from Beijing Jiaotong University (BJTU), Beijing, China, in 2010. During his Ph.D., he visited Media Lab at Nanyang Technological University, Singapore as a joint-Ph.D. student from 2008 to 2010. From 2010 to 2011, he was a Postdoctoral Researcher with the School of Computer Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the Institute of Information Science, Beijing Jiaotong University. His research interests include computer vision, multimedia content analysis, and machine learning.