Unambiguous Scene Text Segmentation With Referring Expression Comprehension

Xuejian Rong, *Student Member, IEEE*, Chucai Yi, *Member, IEEE*, and Yingli Tian, *Fellow, IEEE*

***Abstract*— Text instance provides valuable information for the understanding and interpretation of natural scenes. The rich precise high-level semantics embodied in the text could be beneficial for understanding the world around us, and empower a wide range of real-world applications. While most recent visual phrase grounding approaches focus on general objects, this paper explores extracting designated texts and predicting unambiguous scene text segmentation mask, i.e., scene text segmentation from natural language descriptions (referring expressions) like *orange text on a little boy in black swinging a bat*. The solution of this novel problem enables accurate segmentation of scene text instances from the complex background. In our proposed framework, a unified deep network jointly models visual and linguistic information by encoding both region-level and pixel-level visual features of natural scene images into spatial feature maps, and then decode them into saliency response map of text instances. To conduct quantitative evaluations, we establish a new scene text referring expression segmentation dataset: *COCO-CharRef*. Experimental results demonstrate the effectiveness of the proposed framework on the text instance segmentation task. By combining image-based visual features with language-based textual explanations, our framework outperforms baselines that are derived from state-of-the-art text localization and natural language object retrieval methods on *COCO-CharRef* dataset.**

***Index Terms*— Natural language description, text detection, text retrieval, text recognition, deep neural network, referring expression.**

I. Introduction

AS A category of self-described object entities, text instances such as characters, words, and sentences in a scene image provide one of the most concise and accurate cues to help people understand and interpret natural scenes. Reading text information from scene natural images, namely scene text extraction, could play an important role in image search [1], [2], instant translation [3], robot navigation and self-driving autonomous system [4], industrial automation [5], augmented reality [6], and many other vision-language applications.

Manuscript received May 15, 2019; accepted July 12, 2019. Date of publication July 26, 2019; date of current version September 25, 2019. This work was supported by the National Science Foundation under Award IIS-1400802. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pavan Turaga. *(Corresponding author: Yingli Tian.)*

X. Rong and Y. Tian are with the Department of Electrical Engineering, The City College, The City University of New York, New York, NY 10031 USA (e-mail: xrong@ccny.cuny.edu; ytian@ccny.cuny.edu).

C. Yi was with The Graduate Center, The City University of New York, New York, NY 10031 USA. He is now with Google on Augmented Reality, Mountain View, CA 94043 USA (e-mail: gschucai@gmail.com).

Digital Object Identifier 10.1109/TIP.2019.2930176

Characterness, a.k.a. scene text segmentation mask, as a measure of scene text saliency in natural images first introduced in [7], has been widely used similar to *objectness* for object saliency detection. One main motivation for predicting characterness, a.k.a. scene text segmentation mask, is the fact that *text attracts human attention*, even when amongst a cluttered background [7]. This has been shown by a range of authors including Judd *et al.* [8] and Cerf *et al.* [9] who verified that humans tend to focus on the text in natural scenes. Also, as indicated in Bylinskii *et al.* [10] on the research of saliency prediction and human eye fixations, text instances are widely existing in the daily life and always attract much human visual attention, and quantitatively accounts 29% out of all under-predicted outdoor labels by DeepFix [11] model in the CAT2000 dataset [12].

The *referring expression*-based text saliency and segmentation has several significant benefits for a variety of scene understanding tasks. Compared with a set of unordered text bounding boxes predictions traditionally generated by scene text detection approaches, a precise pixel-wise segmentation of target text instances is more valuable because 1) First, the text bounding boxes of annotated text regions were usually contaminated by background objects, especially for the text instances in irregular shapes. 2) Second, a predicted text bounding box did not completely match its text instance in most cases, and many methods would add post processing to adjust their positions and sizes. These defective text regions would drastically affect the following text recognition, transcription, and other associated tasks depending on the accuracy of text regions. In contrast, segmentation-based pixel-wise text extraction methods could generate more accurate and useful text annotations. 3) Third, accurate pixel-level text segmentation mask is also necessary for many applications such as scene text image inpainting/completion and the automatic removal of image annotations or video subtitles.

However, even though scene text segmentation could provide a more comprehensive extraction output with respect to scene text detection, not all text instances are equivalently important in a scene image. The informativeness of a text instance in the background context, and the allure of the text instance on its own, can affect how long individual observers fixate on it, and what proportion observers gaze at [10]. Most previous approaches of scene text extraction merely regarded text instances as one generic category of objects, and attempted to exhaustively generate their location (*spatial*) and character sequence (*literal*) information while ignoring

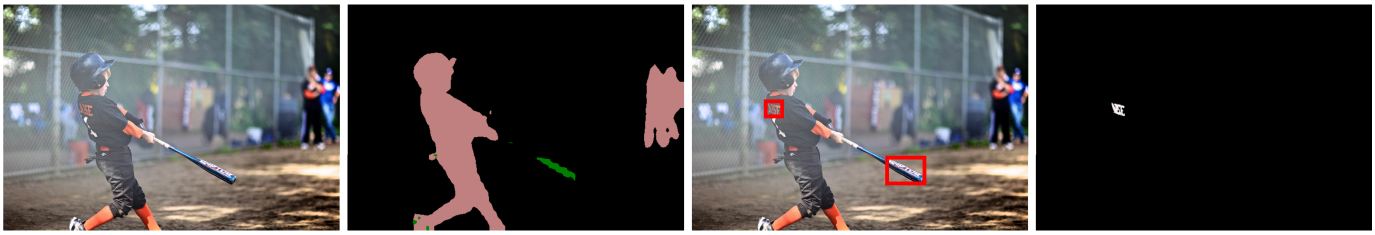

Fig. 1. In this paper we approach the novel problem of *unambiguous scene text segmentation*. Unlike traditional semantic image segmentation and scene text detection/segmentation, the proposed framework jointly models visual features and context descriptions to predict specific scene text segmentation mask, which results in unambiguous pixel-wise scene text segmentation from natural language expressions. From left to right: 1) original image; 2) semantic image segmentation on category *human*; 3) region-level scene text detection represented with bounding boxes; 4) scene text segmentation mask from referring expression query *"orange text on a little boy in black swinging a bat"*.

higher level *semantic* and *contextual* information. It means that text instances were treated the same as other objects for scene understanding and description, even though they are probably straightforward explanations of the surrounding context environment and semantically self-described.

As we know, text information could help people understand surrounding environments only if the relationship between text instances and the context environment were perceived. It means that user-specific text instances could be more valuable than a set of unorganized text instances from scene images. Therefore, this paper addresses the following problem: *from an image containing scene text instances, and a referring expression that describes specific text instances with their context, how to predict the text regions corresponding to the referring expression and further segment text instances in pixel level (*i.e.*, unambiguous scene text segmentation / characterness prediction)*. For example, as shown in Figure 1, for the query *{orange text on a little boy in black swinging a bat}* our framework would predict a segmentation mask that covers the orange word exactly on the target entity, but not the others. This problem is related to but different from recent work on *natural language object retrieval/segmentation* [13], [14], *referring expression comprehension* [15], and *phrase grounding* [16], [17], since the highly variant appearance, scale, and density, and the style of self-description tell the significant difference between text instances and generic objects. Besides, the recent semantic segmentation approaches were not specifically designed for text instances which are usually inconsistent, fragmental, and lack of well-defined shape/boundary and amorphous regions, so they hardly work well on text segmentation. Our proposed model outperforms the state-of-the-art approach on the newly collected benchmark dataset, named COCO-CharRef, for this novel task.

High-quality and user-specific scene text segmentation from referring expressions can underpin many vision-language applications which rely on natural language interfaces, such as communicating with a grocery shopping aid for blind or visually impaired users (e.g., *Alexa, please read me the price of non-fat CHOBANI Greek yogurt on the top shelf*), or interacting with video editing software (e.g., *Premiere, please transcribe/mosaic all the identity information if photo IDs or credit cards appear in this video*). In addition, it is a good testbed for research in the area of vision and language systems on scene text images. The proposed framework could also help dramatically boost the efficiency of the whole scene text reading system while avoiding the exhaustive search and recognition of all text instances.

Contributions: The contributions of our work are three-fold.

- We propose a new framework that effectively segments pixel-level text instances from context description by the prediction of unambiguous scene text segmentation mask.
- The relationship between text instances and their context are explored and modeled in this paper as *{text-predicate-context}* triplets, where the *predicate* can be *spatial, preposition, comparative, visual attributes* or their combinations. While the number of such triplet relationship phrases could grow combinatorially, the proposed model tends to handle a large number of relations by sharing parameters among them. For instance, a single "on" classifier can be learned to recognize both *{text on bags}* and *{text on dogs}*, even when *{text on dogs}* has never been seen in training.
- A new large-scale benchmark dataset is constructed to evaluate the performance of the new task and other text-based scene understanding approaches. And the proposed approach achieves remarkable performance improvement compared with several strong baseline methods.

The remainder of this paper is organized as follows: Sec. II reviews related work recently published. In Sec. III, the baseline methods, and the proposed model are described in detail. The construction of new benchmark dataset, experimental results, and analytic discussions are presented in Sec. IV. We conclude our paper and describe future work in Sec. V.

II. Related Work

Text detection and segmentation, word recognition, word image retrieval, image captioning and description, visual question answering, generation and comprehension of referring expressions, image-language alignment, phrase grounding, and visual relation reasoning/modeling could be considered as diverse subtasks of the supertask of systematic visual and linguistic interaction, which jointly models the natural language and scene image information. We discuss the related work with the proposed new task as follows.

A. Text Extraction in the Wild

Text extraction from the natural scene has been increasingly popular in academic research and practical applications.

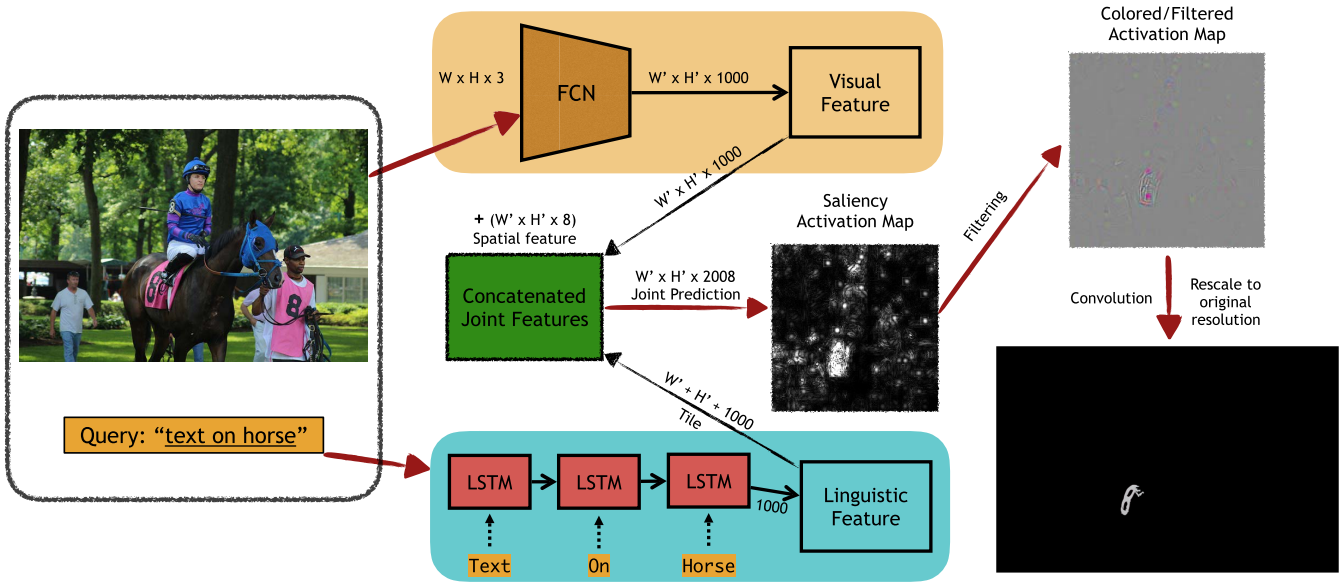

Fig. 2. **The Architecture of Our Proposed Model.** Given an input image containing text instances, the model encodes the image-based data along with the language-based query of referring expression into convolutional and recurrent features, then decodes them into the saliency response map through a fully convolution classification network. After upscaling, the final text segmentation masks are generated from the saliency response map.

Most existing text extraction methods localize scene text instances with word-level bounding box annotations, although they depend on multiple schemes, including sliding-window based [18]–[20], connected components-based [21]–[23], or deep neural network-based [24]–[37]. However, bounding box-based text regions generated by traditional text detection approaches are usually not tight enough and contain many background outliers (especially when text instances are in irregular or curved shapes), which makes the following scene text recognition difficult. In contrast, segmentation-based pixel-wise text extraction methods could generate more accurate and useful text annotations.

In our proposed framework, by modeling both visual and linguistic information in context, text instances are able to be segmented from the scene image with pixel-wise labeling. Most conventional solutions of text segmentation employed a bottom-up pipeline with heuristic features [23], [38], [39] which were not robust and reliable. It results in the high dependency of low-level image filtering. Even though deep neural network could substantially improve generic semantic object segmentation [40]–[42], text segmentation from scene images with the complex background is still an open problem to be addressed. The main challenge is that text instances always reveal high variants of appearances, scales, and structures that are difficult to model. Some recent approaches [30], [34], [43]–[46] based on deep networks presented well-performed text segmentation of zoomed-in views, but could not satisfyingly handle cluttered environments with large-scale and a wide variety of context information from natural scenes. In our proposed framework, text instances and their context are jointly modeled. The relationship modeling of a text instance with its surrounding objects further benefits the characterness prediction and pixel-wise segmentation.

B. Fully Convolutional Networks for Segmentation

Fully Convolutional Networks (FCNs) are deep neural networks consisting of only convolutional (and pooling) layers. FCNs pioneered the use of deep learning for semantic segmentation, and are still one of the most commonly adopted backbones of the state-of-the-art methods for semantic segmentation over a pre-defined set of semantic categories [41], [47], [48]. One advantage of FCNs is that spatial information is consistently preserved, which makes this kind of networks suitable for segmentation tasks that require spatial grid output. In our framework, FCNs are used to extract features and generate segmentation labels, with the purpose of handling image-based text instances at both region and pixel levels.

C. Alignment of Images With Language

Learning correspondences between sentence structure and image regions has been explored with the visual-semantic alignment. This architecture has been used for applications in image retrieval and caption generation [49], [50]. With new datasets proposed which provide bounding box-level natural language annotations [15], recent work has also investigated region-wise image captioning and description for the tasks of natural language object retrieval [13], dense captioning [51], scene graph parsing [52], and visual common sense reasoning [53]. Our proposed framework has a similar idea that aligns a language triplet with regions of pixels in the image. Typically, existing approaches do not explicitly represent relations between noun phrases in a sentence to improve visual-semantic alignment. We believe that understanding these relations will lead to better scene understanding including phrase grounding and comprehension, and scene graph generation and reasoning.

D. Visual Relation Modeling

Triplet learning has been addressed in various tasks such as mining typical relations (knowledge extraction) [54], reasoning [55], object detection [56], or image retrieval [57]. In this work, we address the task of relationship modeling in scene text segmentation from language based explanations. Early work on human-object interactions [58] models the triplet in the form $(person, action, object)$. Recently, the work in [59] tried to generalize the similar setting to non-human subjects by developing a language model sharing knowledge among visual detections related to each other. Inspired by the idea but different from these approaches, we restrict the $subject$ to be a text instance and cover a broader class of predicates that include prepositions and comparatives. In our work this combinatorial challenge can be addressed by developing a new visual representation with better generalization into unseen triplets *{text-predicate-object} and* without depending on a strong language model.

E. Grounding Visual Explanations

Our proposed framework is an innovative combination of the recent work on object localization and segmentation from natural language descriptions, i.e., referring expression comprehension. In those work, the task is to localize/segment a target object in a scene based on its natural language referring expression (by drawing a bounding box over it, or pixel wisely assigning the foreground label to it). The methods of [13] and [15] are built upon image captioning frameworks such as LRCN [60] or mRNN [61], and localize objects by selecting the bounding box where the expression has the highest probability. The authors of [62] firstly proposed a natural language based scene text extraction methods, but the framework is not trained end-to-end and cannot output pixel-wise text annotations. In [16], the authors proposed a model to localize a textual phrase by attending to a region on which the phrase can be best reconstructed. In [63], a joint embedding space of visual features and words is learned to localize target object by searching the closest region in the joint embedding space. Reference [14] proposes an end-to-end training method for generating object segmentation mask from natural language descriptions. The proposed model encodes the given expression into a real-valued vector using LSTM networks [64], and extracts a spatial feature map from the image using a Convolutional Network. Then it performs pixel-wise classification based on the encoded referring expression and feature map to output an image mask covering the visual entity described by the expression. Liu *et al.* [65] further propose to learn the word-to-image interaction instead of modeling image and sentence features independently. The proposed method achieves top results on general object segmentation with language explanations, and also shows that the combination of visual and linguistic features for scene text segmentation is worth exploring.

To the best of our knowledge, all previous methods of natural language based detection and retrieval can only return a bounding box or segmentation mask of non-text generic objects, and no prior work has learned to directly segment text instances given a natural language description as a query. The most related approach with our work is the recent unambiguous text localization and retrieval model (CRTR) proposed in [62], which pioneered the task of natural language based scene text detection and retrieval. However, CRTR is only capable of generating bounding box level prediction, and relies on preceding results from a sequential text detection model. Also, the framework proposed in [62] are not end-to-end trainable (though being end-to-end evaluated). In comparison, we construct strong baselines with the foreground segmentation based on the bounding boxes obtained by combining the state-of-the-art text detection method [27] and natural language retrieval method [13]. The CRTR model proposed in [62] is also regarded as one start-of-the-art to compare with (see details in Sec. IV).

III. Proposed Model

In this section, we first describe our proposed framework in detail from Sec. III-A to Sec.III-C, including text image visual feature encoding, referring expression linguistic feature encoding, fusion feature decoding, and saliency response map upsampling. Then several effective baseline methods are presented, which are derived from previous state-of-the-art approaches in Sec. III-D.

A. Spatial Feature Map Extraction

As shown in Figure 3, given an image with scene text instances, our proposed framework first computes an effective feature representation that is able to encode visual appearance of text characters and their relations with surrounding objects. This feature representation preserve the spatial information to enable the correct spatial prediction of a segmentation mask. This is accomplished by a fully convolutional network model similar to FCN-32s [41], where the image is fed through a series of convolutional (and pooling) layers obtaining a feature map containing encoded spatial information fused from different levels. In our framework, the network is further modified to encode both region-wise and pixel-wise information, resulting in a more effective feature representation compatible with varieties of text instances.

Given an input scene text image of size $W \times H$, we apply a convolutional network to the image and obtain a $W' \times H'$ spatial feature map, with each position on the feature map containing D_{im} channels (D_{im} dimensional local descriptors). At each position of the spatial feature map, the D_{im} dimensional local descriptor is further L2-normalized in order to obtain a more robust feature representation with respect to degradations. In this way, we can extract a $W' \times H' \times D_{im}$ spatial feature map as the representation of each image.

To allow the model to reason about spatial relationships, two extra channels are inserted into the feature maps: the x and y coordinates of each spatial location. The top left corner and the bottom right corner of a feature map are represented as $(-1, -1)$ and $(+1, +1)$, respectively. In this way, we obtain a $w \times h \times (D_{im} + 2)$ feature representation containing both local image descriptors and spatial coordinates.

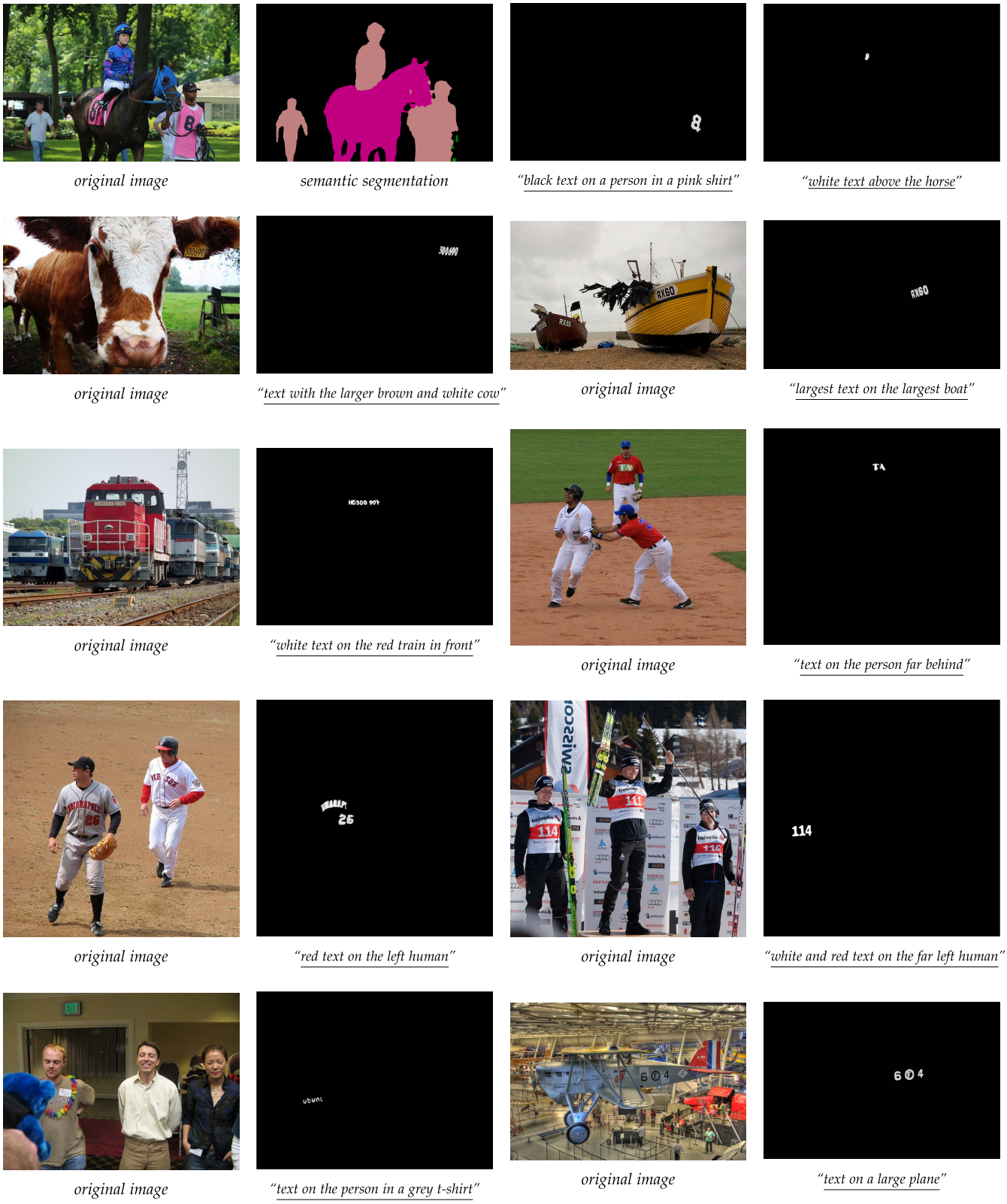

Fig. 3. Examples of unambiguous segmentation of text regions from the *COCO-CharRef* dataset. Text instances in not-high resolutions are often embedded in complex background context and cluttered, but are still successfully segmented accordingly.

In our implementation, the VGG-16 architecture [66] is adopted as a fully convolutional network by transforming fully-connected layers `fc6`, `fc7` and `fc8` to convolutional layers, which outputs $D_{im} = 1000$ dimensional local descriptors. The resulting feature map size is $W' = W/s$ and $H' = H/s$, where $s = 32$ is the pixel stride on fc8 layer

output. It means that a unit on the spatial feature map has a large enough receptive field of 384 pixels, which aggregates the information of context concepts for text instances from neighboring regions. This design can help reason about the interaction between visual text entities and context concepts.

B. Encode Referring Expressions With LSTM

For the input natural language expression that describes a scene text region, we model the query sequence into a vector because it is more efficient to process fixed-length vectors than variable-length sequences. In our encoder, for the natural language expression in a sequence-to-sequence manner [67], each word is first embedded into a vector through a word embedding matrix, and then a recurrent LSTM [64] network with D_{text} dimensional hidden state is used to scan through the embedded word sequence. For a text sequence $S = (w_1, \ldots, w_T)$ with T words (where w_t is the vector embedding for the t-th word), at each iteration t, the LSTM network takes as input the embedded word vector w_t from the word embedding matrix. At the final iteration $t = T$ when the LSTM network has seen the whole text sequence, the hidden state h_T in LSTM network is regarded as the encoded vector representation of the expression. Similar to the encoding of spatial features in Sec. III-A, we also L2-normalize the D_{text} dimensions in h_T and set $D_{text} = 1000$ in our implementation.

Specifically, every word w_t is one-hot encoded and mapped to a word embedding $\mathbf{w}_t$. The entire sentence is then encoded with an LSTM into a vector $\mathbf{h}_T$ of size D_{text}, where $\mathbf{h}_t$ represents the hidden state of LSTM at time step t:

$$\text{LSTM} : (\mathbf{w}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \rightarrow (\mathbf{h}_t, \mathbf{c}_t), \tag{1}$$

$$\mathbf{c}_t = \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{g}, \tag{2}$$

$$\mathbf{h}_t = \mathbf{o} \odot \tanh(\mathbf{c}_t), \tag{3}$$

where n is the size of the LSTM cell. $\mathbf{c}_t$ is the memory states at time step t. The vector $\mathbf{h}_T$ is then concatenated with the image features and spatial coordinates at all locations to produce a $W' \times H' \times (D_{im} + D_{text} + 8)$ tensor.

C. Spatial Classification and Upsampling

After extracting the spatial feature map from a scene image in Sec. III-A and the encoded referring expression h_T in Sec. III-B, we need to determine whether a spatial location on the feature map belongs to the foreground, which denotes the text instances described by the natural language expression. In our framework, this is accomplished by a fully convolutional classifier over the local image descriptor and the encoded referring expression. We first tile and concatenate h_T to the local descriptor at each spatial location in the spatial grid to obtain a $W' \times H' \times D^*$ (where $D^* = D_{im} + D_{text} + 2$) spatial map containing both visual and linguistic features. Then, a two-layer classification network is trained with a D_{cls} dimensional hidden layer, which takes as input the D^* dimensional representation and outputs a score to indicate the probability that a spatial location belongs to the target image region. We set $D_{cls} = 500$ in our implementation.

This classification network contains two 1×1 convolutional layers (with ReLU none linearity between them) and is applied to the underlying $W' \times H'$ feature map. The fully convolutional classification network outputs a $W' \times H'$ coarse *low-resolution scene text response map* containing classification scores, which can be regarded as a low-resolution segmentation of the text instances described by the referring expression.

In order to obtain a segmentation mask with higher resolution, upsampling is further performed by deconvolution (swapping the forward and backward pass of convolution operation) [41], [48]. In our framework a $2s \times 2s$ deconvolution filter is used with stride $s = 32$ for the VGG-16 network architecture, similar to the FCN-32s [41]. The deconvolution filter produces a $W \times H$ *high resolution saliency response map* that has the same size as the input image, and the values on this map represent the confidence of determining if a pixel belongs to target object. We use the pixel-wise classification results (i.e., whether or not a value on the response map is greater than 0) as the final segmentation prediction.

In training phase, a training sample is a tuple (I, T, M), where I is a scene text image, T is a natural language expression describing specific text region(s) within that image, and M is a binary segmentation mask of the corresponding text region(s). The training loss function is defined to be average pixel-wise loss, as presented in the following equation.

$$Loss = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} L(v_{ij}, M_{ij}) \tag{4}$$

where W and H are image width and height, v_{ij} is the response value (score) on the high resolution saliency response map and M_{ij} is the binary ground-truth label at pixel (i, j). L is the per-pixel weighted logistic regression loss as follows

$$L(v_{ij}, M_{ij}) = \begin{cases} \alpha_f \log(1 + \exp(-v_{ij})) & \text{if } M_{ij} = 1 \\ \alpha_b \log(1 + \exp(v_{ij})) & \text{if } M_{ij} = 0 \end{cases} \tag{5}$$

where α_f and α_b are loss weights of foreground and background pixels respectively. In experiments, we observe that training converges faster in the presence of higher loss weights for foreground pixels, and we set $\alpha_f = 3$ and $\alpha_b = 1$ in $L(v_{ij}, M_{ij})$.

All the parameters of the network are initially derived from a VGG-16 network [66], which is pretrained on the 1000-class ILSVRC classification task [68], and fine-tuned on newly collected *COCO-CharRef* dataset. The deconvolution filter for upsampling is initialized from bilinear interpolation. All the other parameters in our proposed framework, including the word embedding matrix, the LSTM parameters and the classifier parameters, are randomly initialized. The whole network is trained with standard back-propagation using SGD with momentum.

D. Baseline Methods

As far as we know, no previous work was able to directly predict the pixel-wise segmentation of text instances based on referring expression in natural language. To evaluate our method, we construct several baseline methods of text segmentation, and compare their performance with our proposed text segmentation framework.

1) Classification Over Characterness *Proposals:* To re-implement this baseline method, we first extract a set of text segmentation proposals using the original Characterness method [7], and then train a binary classifier to determine whether or not a text segmentation proposal matches the expression. First, the encoded query is concatenated with visual features that are extracted from each proposal using a VGG-16 [66] network pretrained on [68]. Then, the concatenated features are adopted to train a classifier to predict a segmentation proposal to be foreground or background. In contrast, our framework employs fully convolutional network to perform pixel-wise classification in an end-to-end way, without relying on the generation of candidate text segmentation.

2) Foreground Segmentation From Text Region Proposals: To re-implement this baseline method, a recent text localization method (TextBoxes [27]) is employed to exhaustively obtain all potential text bounding boxes, which are then scored and ranked by the Spatial Context Recurrent ConvNet (SCRC) [13] based on the given referring expression. The state-of-the-art scene text detector "TextBoxes" [27] is derived from the Single Shot Detector [69]. It achieves highly competitive results on standard text detection benchmarks, f-measure 0.85 on ICDAR 2011/13 dataset. We select the top 50 text region proposals from [27] and input them into SCRC. Next, the foreground segmentation is extracted from the top text bounding boxes using GrabCut [70]. SCRC localizes a referring expression by scoring/ranking candidate bounding boxes based on an image captioning model. Our framework takes from a testing image the top-ranked 50 candidate bounding boxes that best match the associated testing referring expression, and performs two types of evaluations, either using the globally highest scoring bounding box as the segmentation, or the foreground segmentation resulted from GrabCut [70].

IV. Experiments

A. Dataset Construction

Although there are many datasets for the evaluations of scene text segmentation, semantic object segmentation, and image captioning respectively, no benchmark dataset is available with both pixel-level scene text annotations and corresponding natural language descriptions. The ReferIt dataset [71] has been widely used in image captioning and natural language object retrieval. However, it does not provide any image-based annotations or language-based referring expressions for the scene text instances. The ICDAR dataset (Task 2 for Robust Reading Challenges 2013 and 2015) [72] contains real-world images of text on sign boards, books, posters and other objects with pixel-level foreground/background annotations. However, most text bounding boxes from these datasets are in extremely focused view and rarely contain useful context concept entities. Therefore, to evaluate the performance of the baseline methods (as described above in Sec. III) and our proposed framework in the way of unambiguous scene text segmentation, we establish a new *Characterness Referring Expression* dataset named as *COCO-CharRef*. Specifically, it is built on the basis of several existing externally annotated datasets [1] [73]–[75], all of which were further originated from MS COCO [76]

1) MSCOCO Dataset: MSCOCO [76] is an image recognition, segmentation, and captioning dataset. It contains more than $300,000$ images and 2 million instances across 80 object categories. It serves as a fundamental data source for the three datasets introduced below.

2) COCO-Text Dataset: COCO-Text is a large-scale dataset for text detection and recognition in natural images, based on MSCOCO dataset. It contains more than $63,000$ images and $173,000$ text instances. However, it does not provide pixel-level text segmentation annotations and natural language descriptions. Also, there exist many annotated text instances that are illegible for transcription and recognition, resulting from too small text height or too much occlusion. Our dataset ignores these kinds of illegible text samples.

3) RefCOCO Dataset: RefCOCO [75] integrates four earlier referring expression datasets (RefCOCOg, RefCOCO, RefCOCO+, and RefClef), and provides instance level natural language referring expressions. Due to the compensation between COCO-Text and RefCOCO datasets, we select the overlapped part of them to establish half of the new *COCO-CharRef* dataset. The ground truth pixel-level annotations are generated from the ground truth word-level bounding box annotations through GrabCut [70], and further filtered with human assessment. The complete referring expressions for text instances are generated by combining the detected relationship between text instances and context concepts, with the existing referring expressions for those context concepts. The illegible text instances are treated as "Don't Care" regions, and cannot affect the final evaluation performance [72]. In total, we get $2,427$ legible text images for final dataset construction.

4) COCO-Stuff Dataset: COCO-Stuff [74] is a dataset which augments the MSCOCO dataset with pixel-level *stuff* annotations. Besides the originally annotated 80 object classes, COCO-Stuff provides $10,000$ complex images which are further annotated with 91 stuff classes including both indoor-stuff (e.g., floor, wall, ceiling) and outdoor-stuff (e.g., building, ground, sky). To enable the accurate pixel-level evaluation of text segmentation, we follow the text rendering methods in [77] and generate synthetic text images. The complete referring expressions are obtained from the intersection of synthetic COCO-Stuff text images and RefCOCO datasets, which are similar to those in COCO-Text. Since scene text mostly appears in urban artificial environments, in the generation of text images, we discard some categories of outdoor objects that should not belong to those environments. In total $2,573$ text images are generated in the established dataset.

5) New Created COCO-CharRef Dataset: In summary, our constructed COCO-CharRef benchmark dataset is built on the COCO-Text, RefCOCO, and COCO-Stuff datasets. Specifically, it contains $5,000$ images ($4,000$ for training and $1,000$ for evaluation), and $84,112$ referring expressions annotated on $19,241$ text regions. To date, the COCO-CharRef dataset is the biggest available dataset that contains natural language expressions annotated on segmented scene

[1] http://mscoco.org/external/

TABLE I

THE PERFORMANCE OF OUR FRAMEWORK AND BASELINE METHODS ON THE COCO-CHARREF BENCHMARK DATASET UNDER PRECISION METRIC AND OVERALL IOU METRIC. NUMBERS IN BOLD INDICATE THE BEST PERFORMANCE

Method	@0.5	@0.6	@0.7	@0.8	@0.9	overall IoU
Whole image	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%
Characterness classification [7]	3.5%	1.5%	0.4%	0.3%	0.0%	9.4%
SCRC [13] + TextBoxes [27]	5.3%	2.1%	0.7%	0.0%	0.0%	14.9%
SCRC [13] + TextBoxes [27] + GrabCut	7.6%	3.7%	1.9%	0.4%	0.4%	18.2%
TextSeg [14] (retrained)	11.4%	5.1%	2.6%	0.7%	0.3%	20.5%
CRTR [62] + GrabCut	10.7%	5.8%	3.3%	1.8%	0.6%	25.4%
Proposed: low resolution	9.5%	4.2%	2.7%	1.3%	0.4%	24.7%
Proposed: high resolution	**12.3%**	**6.9%**	**4.6%**	**2.1%**	**0.9%**	**28.1%**

text regions as far as we know. The dataset and related evaluation code will be made available.

B. Evaluation on COCO-CharRef Dataset

1) Metrics: On the newly-collected COCO-CharRef dataset, we evaluate the performance of our framework and the baseline methods. Two metrics are used for evaluation: the *overall intersection-over-union* (overall IoU) metric and the *precision* metric. In the experiments, the area of a region is represented by the number of pixels in that region. The overall IoU is the total intersection area divided by the total union area, where both intersection area and union area are accumulated pixel-wisely overall testing samples (notice that a testing sample is a pair of scene image and referring expression as well). Although the overall IoU metric is the standard metric used in PASCAL VOC segmentation [13], our evaluation metric has to measure the accuracy of segmenting foreground text instances specified by the input referring expression from the background. The precision metric goes along with 5 different IoU thresholds from low to high: Precision@X ($X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$) where Precision@X means the percentage of images with IoU higher than X. Precisely, the precision metric is the percentage of test samples where the IoU between prediction and ground-truth passes the threshold. For example, precision@0.5 is the percentage of expressions where the predicted segmentation overlaps with the ground-truth text region by at least 50% IoU.

2) Results: The evaluation results of our experiments are summarized in Table I. By simply returning the whole image, the baseline method only achieves 0.7% overall IoU. This is partially caused by the fact that COCO-CharRef dataset contains some large context regions and the overall IoU metric put more weights on large regions such as walls and buildings. Moreover, a reasonable overall IoU can be obtained through per-word segmentation (Per-word). The prediction of text bounding box as a whole from SCRC [13] (`SCRC + TextBoxes`) obtains better performance than the prediction of the referring expression-based segmentation proposal [7] ("Characterness classification"). Also, it shows that GrabCut [70] obtains better performance of foreground segmentation from SCRC-ranked TextBoxes (`SCRC + TextBoxes + GrabCut`) than just (`SCRC + TextBoxes`).

As to the method of end-to-end object segmentation from language descriptions proposed in [14], we retrain the segmentation network on scene text datasets [72] and conduct the evaluation on the COCO-CharRef dataset. The state-of-the-art referring scene text localization method is also combined with GrabCut with the performance presented. In summary, [14] achieves decent precision on lower settings, but does not perform well on higher settings (P@0.8 and P@0.9). The combined {CRTR + GrabCut} model achieves decent and consistent performance, but still inferior compared with the high-resolution version of the proposed model.

We believe that the Precision metric is more compatible with the task of text instance segmentation from natural language descriptions. In practical applications, the user would prefer the specified segmentation from their referring expression query, rather than the detection or segmentation accuracy of all text instances.

As the evaluation results, our proposed framework outperforms all the competing methods by a large margin under both precision metric and overall IoU metric. In Table I, the second last row ("low resolution") denotes the process of bi-linear upsampling over the coarse response map obtained from the coarse-level model in low-resolution space, and the last row ("high resolution") denotes the performance of the full model in high-resolution space, and the effectiveness of spatial upsampling at low-resolution space. In addition, it demonstrates that the final high-resolution model achieves both higher precision and higher overall IoU, compared with the baseline methods. Some examples of text instance segmentation are illustrated in Figure 3, where the top row depicts different segmentation results from different referring expressions on the same image. Images and other rows present more unambiguous text segmentation results.

Figure 4 illustrates some failure cases on COCO-CharRef dataset, where the IoU between prediction and ground-truth segmentation is less than 50%. In some examples, our framework fails to accurately segment the strokes of text

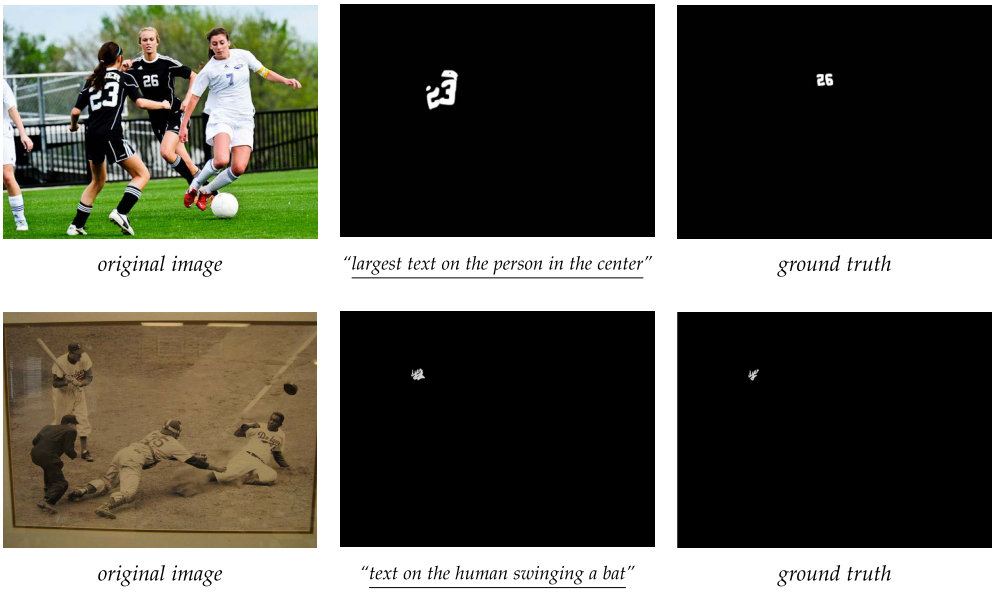

Fig. 4. Some failure cases where IoU < 50% between prediction and ground-truth, due to incorrect region-level prediction or inaccurate pixel-level estimation of the corresponding models.

instances, but still produces reasonable text saliency response maps covering the target regions specified by natural language referring expressions. Our framework could be continuously fine-tuned by integrating more training examples associated with the failure cases.

V. Conclusion

In this paper, we have addressed the challenging problem of unambiguous scene text segmentation, i.e. segmenting specific scene text instances described by the referring expressions from natural scenes. In our proposed framework, spatial information and context descriptions of scene text instances benefit from each other. Scene text instances could provide pivotal and precise information for context descriptions of the entire or a region of the scene image, while context description could provide a more user-friendly way to incorporate the extracted text information and its context into practical applications. Experimental results on the newly collected benchmark dataset demonstrate that our framework outperforms baseline methods by a large margin. Our future work will focus on more comprehensive modeling of the relationships between scene text instances and their contexts.

References

[1] G. Schroth, S. Hilsenbeck, R. Huitl, F. Schweiger, and E. Steinbach, "Exploiting text-related features for content-based image retrieval," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2011, pp. 77–84.

[2] S. S. Tsai, H. Chen, D. Chen, G. Schroth, R. Grzeszczuk, and B. Girod, "Mobile visual search on printed documents using text and low bit-rate features," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2601–2604.

[3] C. Parkinson, J. J. Jacobsen, D. B. Ferguson, and S. A. Pombo, "Instant translation system," U.S. Patent 9 507 772 B2, Nov. 29, 2016.

[4] R. Schulz *et al.*, "Robot navigation using human cues: A robot navigation system for symbolic goal-directed exploration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1100–1105.

[5] Z. He, J. Liu, H. Ma, and P. Li, "A new automatic extraction method of container identity codes," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 1, pp. 72–78, Mar. 2005.

[6] M. Petter, V. Fragoso, M. Turk, and C. Baur, "Automatic text detection for mobile augmented reality translation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 48–55.

[7] Y. Li, W. Jia, C. Shen, and A. van den Hengel, "Characterness: An indicator of text in the wild," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1666–1677, 2014.

[8] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.

[9] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, p. 10, 2009.

[10] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 809–824.

[11] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4446–4456, Sep. 2017.

[12] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," in *Proc. CVPR Workshop*, 2015, pp. 1–4.

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

[14] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 108–124.

[15] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 11–20.

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

[17] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng, "Structured matching for phrase localization," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 696–711.

[18] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust text detection in natural scene images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 970–983, May 2014.

[19] Q. Ye and D. Doermann, "Text detection and recognition in imagery: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1480–1500, Jul. 2015.

[20] L. Neumann and J. Matas, "Efficient scene text localization and recognition with local character refinement," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 746–750.

[21] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.

[22] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao, "Multi-orientation scene text detection with adaptive clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1930–1937, Sep. 2015.

[23] L. Neumann and J. Matas, "Real-time lexicon-free scene text localization and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1872–1885, Sep. 2016.

[24] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in *Proc. CVPR*, Jun. 2016, pp. 4159–4167.

[25] Z. Tian, W. Huang, T. He, Pan. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. ECCV*, 2016, pp. 56–72.

[26] T. He, W. Huang, Y. Qiao, and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," Mar. 2016, *arXiv:1603.09423*. [Online]. Available: https://arxiv.org/abs/1603.09423

[27] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A fast text detector with a single deep neural network," in *Proc. AAAI*, 2017, pp. 4161–4167.

[28] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 3047–3055.

[29] Y. Liu and L. Jin, "Deep matching prior network: Toward tighter multi-oriented text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1962–1969.

[30] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2550–2558.

[31] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5551–5560.

[32] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, "Detecting curve text in the wild: New dataset and new solution," 2017, *arXiv:1712.02170*. [Online]. Available: https://arxiv.org/abs/1712.02170

[33] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: https://arxiv.org/abs/1706.09579

[34] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting scene text via instance segmentation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[35] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5020–5029.

[36] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, "FOTS: Fast oriented text spotting with a unified network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5676–5685.

[37] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jun. 2018, pp. 67–83.

[38] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3538–3545.

[39] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan, "Scene text extraction based on edges and support vector regression," *Int. J. Document Anal. Recognit.*, vol. 18, no. 2, pp. 125–135, 2015.

[40] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

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

[42] P. Wang *et al.*, "Understanding convolution for semantic segmentation," 2017, *arXiv:1702.08502*. [Online]. Available: https://arxiv.org/abs/1702.08502

[43] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 512–528.

[44] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.

[45] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1509–1520, Mar. 2017.

[46] J. Fabrizio, M. Robert-Seidowsky, S. Dubuisson, S. Calarasanu, and R. Boissel, "TextCatcher: A method to detect curved and challenging text in natural scenes," *Int. J. Document Anal. Recognit.*, vol. 19, no. 2, pp. 99–117, 2016.

[47] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1529–1537.

[48] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1520–1528.

[49] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[50] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[51] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4565–4574.

[52] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[53] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," 2018, *arXiv:1811.10830*. [Online]. Available: https://arxiv.org/abs/1811.10830

[54] M. Yatskar, V. Ordonez, and A. Farhadi, "Stating the obvious: Extracting visual common sense knowledge," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 193–198.

[55] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 926–934.

[56] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1745–1752.

[57] J. Johnson *et al.*, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3668–3678.

[58] V. Ramanathan *et al.*, "Learning semantic relationships for better action retrieval in images," in *Proc. CVPR*, 2015, pp. 1100–1109.

[59] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 852–869.

[60] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, 2015, pp. 2625–2634.

[61] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*. [Online]. Available: https://arxiv.org/abs/1412.6632

[62] X. Rong, C. Yi, and Y. Tian, "Unambiguous text localization and retrieval for cluttered scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3279–3287.

[63] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.

[64] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[65] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," 2017, *arXiv:1703.07939*. [Online]. Available: https://arxiv.org/abs/1703.07939

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[67] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[68] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

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

[70] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

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

[72] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2013, pp. 1484–1493.

[73] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-text: Dataset and benchmark for text detection and recognition in natural images," 2016, *arXiv:1601.07140*. [Online]. Available: https://arxiv.org/abs/1601.07140
[74] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-stuff: Thing and stuff classes in context," 2017, *arXiv:1612.03716*. [Online]. Available: https://arxiv.org/abs/1612.03716
[75] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 69–85.
[76] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
[77] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. CVPR*, 2016, pp. 2315–2324.

Xuejian Rong (S'14) received the B.E. degree (Hons.) from the Nanjing University of Aeronautics and Astronautics in 2013. He is currently pursuing the Ph.D. degree advised by Prof. Y. Tian with the City University of New York. He is currently a Researcher with Media Lab of The City College, The City University of New York, where he works in the intersection of deep learning, computer vision, and computational photography. His Ph.D. research broadly covers the scene understanding area, specifically on scene text extraction (including scene text detection, retrieval, and recognition), i.e., how to localize and transcribe useful text instances (e.g., characters, words, and sentences) widely existing in our surrounding environments. The related methods can be used in many applications, including image retrieval, instant translation, robots navigation, and PhotoOCR. He also occasionally works on image restoration and other inverse problems in computer vision.

Chucai Yi (S'12–M'14) received the Ph.D. degree in computer science from The Graduate Center of City University of New York, in 2014. Before joining Google, he was with HERE Map Technologies and Amazon Corporates. He is currently with Google Corporate on Computer Vision Research and Development. As a Professional Software Developer, he has been designing and implementing end-to-end applications and services that apply machine learning and computer vision techniques to solve specific problems on large-scale data. As an Academic Researcher, his research work includes object/signage detection and recognition, human involved event detection, 3D point cloud processing, and camera calibration.

Yingli Tian (M'99–SM'01–F'18) received the B.S. and M.S. degrees from Tianjin University, China, in 1987 and 1990, respectively, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 1996. After holding a Faculty position with the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing. She joined Carnegie Mellon University in 1998, where she was a Postdoctoral Fellow with the Robotics Institute. She then worked as a Research Staff Member in IBM T. J. Watson Research Center, from 2001 to 2008. She is one of the Inventors of the IBM Smart Surveillance Solutions. She has been a Professor with the Department of Electrical Engineering, The City College and the Department of Computer Science, The Graduate Center, The City University of New York, since 2008. Her current research focuses on a wide range of computer vision problems from scene understanding, human behavior analysis, facial expression recognition to assistive technology, and medical imaging analysis.