GISCA: Gradient-Inductive Segmentation Network With Contextual Attention for Scene Text Detection

MENG CAO¹, YUEXIAN ZOU¹,², DONGMING YANG¹, AND CHAO LIU¹

¹ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China
²Peng Cheng Laboratory, Shenzhen 518055, China

Corresponding author: Yuexian Zou (zouyx@pkusz.edu.cn)

This work was supported in part by the Shenzhen Science and Technology Fundamental Research Program under Grant JCYJ20160330095814461, and in part by the National Engineering Laboratory for Video Technology–Shenzhen Division and Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality under Grant ZDSYS201703031405467.

ABSTRACT Scene text detection (STD) is an irreplaceable step in a scene text reading system. It remains a more challenging task than general object detection since text objects are of arbitrary orientations and varying sizes. Generally, segmentation methods that use U-Net or hourglass-like networks are the mainstream approaches in multi-oriented text detection tasks. However, experience has shown that text-like objects in the complex background have high response values on the output feature map of U-Net, which leads to the severe false positive detection rate and degrades the STD performance. To tackle this issue, an adaptive soft attention mechanism called contextual attention module (CAM) is devised to integrate into U-Net to highlight salient areas and meanwhile retains more detail information. Besides, the gradient vanishing and exploding problems make U-Net more difficult to train because of the nonlinear deconvolution layer used in the up-sampling process. To facilitate the training process, a gradient-inductive module (GIM) is carefully designed to provide a linear bypass to make the gradient back-propagation process more stable. Accordingly, an end-to-end trainable Gradient-Inductive Segmentation network with Contextual Attention is proposed (GISCA). The experimental results on three public benchmarks have demonstrated that the proposed GISCA achieves the state-of-the-art results in terms of \textit{f-measure}: 92.1\%, 87.3\%, and 81.4\% for ICDAR 2013, ICDAR 2015, and MSRA TD500, respectively.

INDEX TERMS Scene text detection, multi-oriented text, segmentation network, contextual attention, gradient vanishing/exploding problems.

I. INTRODUCTION

Scene text is one of the most common objects in the nature, which frequently appears on many practical scenes and contains important information for many applications [1]–[4], such as blind navigation, scene understanding, autonomous driving, etc. Scene text reading is thus of great importance in computer vision which has made tremendous progress benefiting from the development of Deep Convolution Neural Networks (DCNNs) [5]. Generally, scene text reading can be divided into two main sub-tasks: scene text detection (STD) and scene text recognition. Despite the similarity to traditional OCR, it is noticed that STD is still challenging because text instances exhibit vast diversity in fonts, scales, arbitrary orientations, as well as uncontrollable illumination affects [6], etc.

Thanks to the recent development of general object detection [7], [8] and instance segmentation methods [9], STD has witnessed a gigantic progress. Deep learning based methods for STD directly learn hierarchical features from input images, which demonstrates more accurate and efficient performance in various STD public benchmarks. However, due to the characteristics of scene text, there still remain many problems to address in practice. Firstly, false positive (FP) is a common mistake in STD, which means detectors tend to mistake non-text areas for text areas. Some examples are given in Figure 1. Thus, the FP problem may significantly affect
VOLUME 7, 2019

CAM not only suppresses the background clutter but also The attention mechanism is widely adopted in various methods. This has become a bottleneck to improve STD performance. Therefore, designing a suitable backbone network is crucial to boost the STD performance.

To solve these two problems, we propose a Gradient-Inductive Segmentation network with Contextual Attention named as GISCA for multi-oriented scene text detection. The backbone network of GISCA incorporates two innovative modules, namely Contextual Attention Module (CAM) and Gradient-inductive Module (GIM) to solve the two problems mentioned above respectively. Concretely, CAMs adaptively make the network focus on the salient areas and suppress the background information, thus alleviating the FP problem. As for GIM, this well-designed module has a linear bypass which makes the gradient back-propagation process more stable. To the best of our knowledge, it is the first time to propose this kind of linear/non-linear decouple mechanism in the STD task. Following the convention of text segmentation methods, this novel STD detector generates the detection results directly from the instance segmentation map. Besides, it’s worth mentioning that the segmentation map is generated through two kinds of pixel-wise prediction, text/non-text prediction and link prediction. Similar to PixelLink [11], the link predictions between two pixels are labeled as positive if they are within the same text instance, otherwise they are labeled as negative. Therefore, the text score map and the eight link maps (every pixel has 8 neighbors) form the final segmentation map using the positive linkage to link the positive pixels. The final bounding boxes are generated by applying minAreaRect method in OpenCV to the text instances. With these delicate designs, GISCA reveals a competitive performance in STD. We will show by experiments in Section IV that our proposed detector outperforms other state-of-the-art methods.

The baseline model of this study called PixelLink was proposed in [11]. Our proposed method extends PixelLink with two delicately designed modules CAM and GIM. The attention mechanism is widely adopted in various computer vision tasks. Compared to previous methods, CAM not only suppresses the background clutter but also enhances the detail expression of multi-layer feature maps. As for the gradient back-propagation problems, this paper extensively studies related publications and proposes the Gradient-inductive Module which provides an effective linear gradient back-propagation bypass during the up-sampling process.

The main contributions of this paper are summarized as follows: 1) This paper proposes an adaptive soft Contextual Attention Module (CAM) which serves as a probability heat map and efficiently highlights the salient areas. Compared to other attention mechanisms, CAM avoids the additional input supervisory constraints and skillful loss function designs. Experiments show CAM efficiently extracts meaningful information from the input feature maps and enhances the detail characteristics in salient areas which make the network more suitable for pixel-level and region-level tasks. 2) A novel up-sampling pattern named Gradient-inductive Module (GIM) is proposed to provide both sufficient feature expression and convenient gradient back propagation. GIM facilitates the training process and improves the detection results at the same time. 3) The proposed GISCA outperforms state-of-the-art algorithms on three widely adopted multi-oriented text datasets.

The rest of the paper is organized as follows. Section II reviews some related works of our study. Section III presents pipelines and algorithms of our proposed GISCA. Section IV presents the extensive experimental results. Finally, section V draws the conclusion.

II. RELATED WORK

As a special case of general object detection, scene text detection (STD) basically follows the same framework of objection detection. In section A, we will first review the mainstream approaches in the development progress of scene text detection. Then issues about the application of attention mechanisms are analyzed in Section B. Finally, Section C describes the methods proposed recently for alleviating the gradient vanishing/exploding problems.

A. SCENE TEXT DETECTION

STD aims to localize text instances in images commonly using word bounding boxes. Primarily, most recent methods may roughly be classified into regression-based and segmentation-based methods.

1) REGRESSION-BASED METHODS

More specifically, regression-based methods can be further divided into two categories according to the regression target: proposal-based and part-based methods.

Proposal-based methods follow the routine of one-stage or two-stage detectors such as SSD [8] and Faster-RCNN [7]. TextBoxes [13] modifies SSD [8] for scene text detection with long default boxes and the vertical offset strategy to deal with extreme aspect ratio texts. TextBoxes++ [14] extends TextBoxes by using quadrilateral bounding boxes to detect multi-oriented texts. RRPN [15] transforms the RPN in Faster
RCNN to Rotated Region Proposal Network which facilitates the detection of multi-oriented texts. Wordsup [16] takes notice that there are few character level annotations in STD and introduces the weak supervision strategy to adjust character coordinates. EAST [17] eliminates the steps of candidate aggregation and word partition. Instead, it predicts words or text lines of arbitrary orientations and quadrilateral shapes directly. Liao et al. propose to separate text/non-text classification and regression tasks using rotation-invariant and sensitive features [18]. Wang et al. come up with an Instance Transformation Network (ITN) [19] to learn geometry-aware representation tailored for scene text in an end-to-end network.

Part-based methods tend to regress text parts and predict the linking relationships between them. CTPN [20] proposes a vertical anchor mechanism to predict the location and text/non-text scores for each proposal. Then a recurrent neural network (RNN) is utilized to connect the sequential proposals. SegLink [21] decomposes text into two elements: segments which are bounding boxes covering part of the word and links which connect two adjacent segments. MCN [22] converts an image into a Stochastic Flow Graph (SFG) and then conducts Markov Clustering on this graph to form the final bounding boxes.

2) SEGMENTATION-BASED METHOD
Segmentation-based methods regard the detection task as the instance segmentation problem. Multi-oriented text detection is especially suitable to follow the segmentation method although it invokes complex post-processing. Zhang et al. apply FCN to predict the salient map of text region followed by the character component combination [23]. HMCP [24] utilizes the modified HED to obtain the text region map, character map and linking orientation map. Then the text line results are achieved through the problem-specific graph model. He et al. employ a novel text detection algorithm using two cascaded FCN, one to extract text block regions and the other to remove false positive [25]. Wu et al. develop a lightweight FCN to effectively learn the borders in text images and introduce the border class to text detection [26].

Considering the superiority of the segmentation-based methods in detecting multi-oriented scene texts, this paper uses segmentation as the baseline route.

B. ATTENTION MECHANISM
An important feature of human perception is that a person does not process the entire scene at once when parsing the picture. Instead, humans selectively focus on a certain part of the visual space to get the salient information which presents an overall understanding. This well-known attention mechanism is widely studied in various computer vision tasks. Generally, the attention mechanism falls under two main categories: hard attention [27] and soft attention [28], [29]. Hard attention, e.g. iterative region proposal and cropping, is often a fixed correction mechanism and it tends to be non-differentiable, which makes model training more difficult. In contrast, the trainable soft attention mechanism is a more simple and efficient way to highlight important areas without introducing additional inputs. More importantly, trainable soft attention places only a small computation burden on the model.

Attention mechanisms are widely adopted in scene text detection (STD) to improve the detection accuracy. Text-attention CNN [30] integrates a mask branch in the network and thus provides rich supervision information. This mask branch enhances the capability to discriminate ambiguous text instances. SSTD [31] takes the pixel-wise binary mask of text images as an additional input, which is used to refine the original feature map. Huang et al. incorporate the proposed Pyramid Attention Network (PAN) into the Mask R-CNN framework to enhance the text detection performance [32]. $R^2$CNN++ [33] adopts a multi-dimensional attention mechanism which contains both pixel-wise and channel-wise attention to reduce the adverse impact of noise.

Most of the methods mentioned above either use additional input or require the problem-specific loss function to constrain the optimization process. In this paper, the proposed CAM only uses the contextual information of the network and does not require explicit supervisory constraints. More importantly, CAM not only suppresses background clutter, but also enriches the detail features of salient regions.

C. GRADIENT VANISH PROBLEMS
DCNNs tend to be substantially deep to achieve accurate results, which has been proved in [10]. However, many recent publications [10], [34]–[37] address the problem that the gradient update information may vanish or explode during the back-propagation process. Almost all papers that deal with gradient vanish/explosion problems design a shortcut connection between layers close to the input and those close to the output. More specifically, the shortcut connections are implemented either by residual blocks or by concatenation operations.

For residual block based models, they utilize the residual learning framework proposed in [10] to ease the network training process and the following degradation problem [10]. ResNet [10] hypothesizes that the residual functions are easier to optimize than the original unknown mapping functions. Therefore, ResNet proposes identity shortcuts and reformulates the layers to learn the residual functions. Inspired by Long Short-Term Memory recurrent networks (LSTM), Highway Network [34] designs the bypassing paths and the gating units which provide a method to effectively train networks with hundreds of layers in an end-to-end manner. Stochastic Depth [35] randomly drops residual blocks during the training process to further mitigate the gradient vanishing problem.

The concatenation operation has also been extensively used as another way for alleviating gradient vanishing/exploding problems. DenseNet [36] connects each layer to every other layers in the proposed Dense Block. Every two
FIGURE 2. Architecture of GISCA. Firstly, An image is fed to the down-sampling VGG16 process, namely the $D_1$–$D_5$ layers. $D_5$ is generated by converting fc6/fc7 to the convolution layer using $1 \times 1$ filter. The annotations in parentheses are the same as those defined in [12]. Note that #c stands for the number of channels. #c: 2/16 stands for feature maps with 2 or 16 channels, for text/non-text prediction or link prediction individually. Then $D_5$ is input to the deconvolution layer and $U_4$ is the output. The red dotted box contains two separate modules CAM and GIM. The attention signal is refined through CAM (the grey block depicted in the picture). After the bilinear interpolation, the previous up-sampling layer and the corresponding down-sampling layer are input into GIM to generate the output of the next layer. The final predictions (text/non-text prediction and link prediction) are conducted on $U_1$ which maintains the same resolution as the input image. The eight feature maps in blue dashed box stand for the link predictions in eight directions. The post-processing procedure is adopted by connecting positive pixels with positive links.

Dense blocks are connected by transition layers to change feature map sizes via convolution and pooling. Hyper-column network [37] defines the pixel-level hypercolumn as the vector of activation values of all the shallow layers in CNN, which proves to perform well in fine-grained localization tasks. For the segmentation task, U-Net [38] designs a skip connection to concatenate the down-sampling and up-sampling feature maps.

However, when carefully evaluating these structures, we find that they cannot achieve the direct gradient back-propagation in the whole training process. For example, residual Block based methods have to utilize $1 \times 1$ convolution layers in the bypass branches to deal with the channel mismatch between input and output layers and this non-linear bypass is unfavorable for gradient back-propagation. For DenseNet, the dense concatenation is limited in the dense block because only feature maps in the same dense block preserve the same resolution while the resize of feature map is implemented through convolution and pooling operations. Therefore, only the gradient within the DenseBlock is preserved and the gradient back-propagation between the DenseBlocks is still stuck. U-Net does connect the feature maps between up-sampling and down-sampling processes, but the up-sampling process is still implemented through the non-linear deconvolution operation. To achieve the direct gradient back-propagation, this paper delicately designs a novel Gradient-inductive Module which contains a linear bypass in the up-sampling process. With our design, GIM enables the gradient from the very deep layers to be directly propagated to shallow layers.

III. METHODOLOGY

A. OVERVIEW

The proposed GISCA is an end-to-end trainable, fully convolutional network to detect multi-oriented texts. The overall architecture is presented in Figure 2. The framework can be
Algorithm 1 Algorithm for the Proposed Method

**Input:** RGB images $D_0$

**Output:** detection results

1. **Down-sampling process**
   a. for $i \in [1, 5]$
   b. $D_i = pooling(ReLU(Conv(D_{i-1})))$
2. **Up-sampling process**
   a. if ($i == 4$)
   b. $U_i = deconv(D_5)$
3. **Pixel-level prediction process**
   a. if ($i \in [1, 3]$)
   b. $D_i^j = CAM(D_i, U_i^{i+1})$
   c. $U_i = GIM(D_i, U_i^{i+1})$

4. **Post-processing process**
   a. The positive pixels are grouped with the positive linkage.

The false positives (FP) problem remains a big challenge for scene text detection [39]. Because of the complex background interference and the large aspect ratio of the text shape, the current text detection algorithms tend to mistake some similar objects such as a batch of fence or certain disc-shaped objects as the text areas. The cause of misclassification is probably lack of local detail information in the deep feature maps. Generally, the detail information will be filtered during the convolutional down-sampling process. One typical down-sampling process is depicted in Figure 3. From Figure 3, we can see that low-level feature maps capture more details but they cannot clearly indicate where the text areas are. On the opposite, the high-level feature maps explicitly demonstrate the text areas with the cost of detail information loss.

Although the up-sampling process of U-Net outputs the feature map with the same resolution as the input image, the detail information is still missing during the tedious convolution proceeding. As a result, these feature maps are not suitable for pixel-level or region-level tasks owing to the contradiction between the resolution and semantic information. Therefore, it is important for the trainable model to focus on the salient areas on one hand and retain more detail information on the other hand.

Based on the soft attention mechanism, we propose the Contextual Attention Module (CAM). The architecture of CAM is illustrated in Figure 4, from which we can see that the output feature map of CAM retains both detail characteristics and high-level semantic information (highlighted areas indicating where texts are). The common skip connections in U-Net are replaced by CAMs during the up-sampling process. Compared to the traditional U-Net models, CAM adaptively emphasizes the salient areas and suppresses irrelevant background information without adding a large number of model parameters. Figure 5 illustrates the input feature maps and the output refinement results of CAM. We can clearly see that the glass curtain walls in the input feature maps have high response values, while the CAM output feature maps are more focused on the text area.

The attention map has shape $W_{in} \times H_{in} \times C_{in}$ with each attention coefficient ranging from 0 to 1. The output of CAM is the element-wise multiplication result of the original skip connection feature map (from low-level layers) and the attention coefficient map. As depicted in Figure 4, CAM takes the feature map $D_i$ and $U_{i+1}, i = 1, 2, 3$ as inputs. The adaptive attention map is generated through two $3 \times 3$ convolutional layers and one $1 \times 1$ convolutional layer. The generated attention map has two channel representing text/non-text scores respectively. Therefore, the designed generation process of CAM is as follows:

$$m = ReLU(Conv_{3x3}(D_i) + Conv_{3x3}(U_{i+1}))$$

(1)
FIGURE 3. The illustration of down-sampling process results. (a) The input image. (b)–(f) The feature map $D_1$–$D_5$. Lower level feature maps capture more local details while higher level feature maps contain more semantic information.

FIGURE 4. The architecture of CAM. $D_i, i = 1, 2, 3,$ represents the feature map extracted from the down-sampling process. $U_{i+1}'$ denotes the bilinear interpolation of $U_{i+1}$. The output result $D_i'$ contains both the details in $D_i$ and the semantic information in $U_{i+1}'$.

FIGURE 5. Comparisons of the (a) input and (b) output of CAM. The symbol in the upper right corner is the name of the feature map.

In both Eq(1) and Eq(2), the rectified linear unit $ReLU$ is utilized as the activation function. The final attention result is given as follows:

$$D_i' = s^* \odot D_i$$

(3)

where $s^*$ is generated by broadcasting $s$ to the same channel as $D_i$, namely $C_{in}$. $\odot$ represents the pixel-to-pixel multiplication of $s^*$ and $D_i$.

C. GRADIENT-INDUCTIVE MODULE (GIM)

Generally, for the segmentation task, the well-known networks such U-Net [38], FPN [40] and stacked hourglass [41] having the hourglass-like structure are widely used to obtain deep feature maps with high resolutions. Unfortunately, the training process of the networks which consist of both down-sampling and up-sampling processes becomes more difficult when the DCNNs become increasingly deeper. This phenomenon can be attributed to the gradient vanishing [42] or gradient exploding problems [43] when the gradients propagate from deep feature maps to shallow ones. For instance, the parameter update may stuck in a local optimal region or the gradient value may become infinite.

To address this problem, we carefully design a novel Gradient-inductive Module (GIM), which enables the gradient update information to be directly propagated to shallow layers. Considering that the nonlinear structure is deft at formulating the mapping between input and output while the linear structure can easily propagate gradients, GIM is

\[ S = e^{Softmax(ReLU(Conv_{3\times3}(m)))} \]  

(2)

where $Conv_{3\times3}$ denotes the convolution layer with kernel size $3\times3$ and $Conv_{1\times1}$ represents $1\times1$ convolution. $Softmax$ function is used to normalize the attention coefficients. Through the exponential function, the salient map is enhanced because the gap between text and non-text areas becomes larger.
The concatenation result of $D_i'$ and $U_{i+1}'$ is defined as $M_i$ in Eq (5). The convolution operations are utilized to extract the messages from $M_i$ and adjust the number of channels, as described in Eq (6). With our design, the nonlinear branch of GIM outputs the feature map with the same shape as the input. As shown in equation (7), the linear branch of GIM takes the element-wise addition results of $D_i'$ and $\tilde{M}_i$ as the final outputs. In this manner, feature maps $U_3$, $U_2$, $U_1$ are generated sequentially.

The explanation of GIM’s principle may be given from the perspective of mathematical derivation. Ignoring the influence of CAM, we only integrate GIM in U-Net. According to the architecture of the proposed backbone network shown in Figure 2, we have the following expression:

$$U_i = D_i + \mathcal{H}_\phi(D_i, U_{i+1}), \quad i = 1, 2, 3$$

(8)

where $\mathcal{H}_\phi$ is the asymptotically approximation function in the up-sampling process. Sequentially from Eq (8), we get the following equation:

$$U_1 = D_1 + D_2 + D_3 + \sum_{i=1}^{3} \mathcal{H}_\phi(D_i, U_{i+1})$$

(9)

For convenience, $\mathcal{H}_\phi$ is adopted here to represent the simplification result of nested iterations of $\mathcal{H}_\phi$, $i = 1, 2, 3$. Using the chain rule, we can derive as follows:

$$\frac{\partial \mathcal{L}}{\partial D_1} = \frac{\partial \mathcal{L}}{\partial U_1} \cdot \frac{\partial U_1}{\partial D_1}$$

$$= \frac{\partial \mathcal{L}}{\partial U_1} (1 + \frac{\partial}{\partial U_1} \sum_{i=1}^{3} \mathcal{F}(D_i, U_{i+1})), \quad i = 1, 2, 3$$

(10)

where $\mathcal{L}$ is the loss function. The additive term $\frac{\partial \mathcal{L}}{\partial U_1}$ ensures that the gradient of deep feature map $U_1$ can be directly propagated to the shallow one $D_1$. Therefore, the Gradient-inductive Module (GIM) provides a bypass which “induce” the deep gradient to the shallow feature map. It comes out to be efficient to alleviate the gradient vanishing problem and that is why GIM is named as Gradient-inductive Module. Similarly, we find that the gradient of $U_2$ and $U_3$ can be propagated to $D_2$ and $D_3$ respectively.

$$\frac{\partial \mathcal{L}}{\partial D_2} = \frac{\partial \mathcal{L}}{\partial U_2} (1 + \sum_{i=2}^{3} \frac{\partial}{\partial U_2} \mathcal{G}(D_i, U_{i+1}))$$

(11)

$$\frac{\partial \mathcal{L}}{\partial D_3} = \frac{\partial \mathcal{L}}{\partial U_3} (1 + \sum_{i=2}^{3} \frac{\partial}{\partial U_3} \mathcal{M}(D_i, U_{i+1}))$$

(12)

where $\mathcal{G}$ and $\mathcal{M}$ are defined as the nonlinear mapping function. It is noted that the existence of $\frac{\partial \mathcal{L}}{\partial U_2}$ and $\frac{\partial \mathcal{L}}{\partial U_3}$ makes the gradient update in layers $D_2$ and $D_3$ more stable.

**D. PIXEL LEVEL PREDICTION**

The segmentation method is widely used in STD by casting the detection task as text/non-text classification tasks especially when dealing with the multi-oriented or arbitrary-oriented text detection. This paper follows the

![Flowchart of GIM](image)

**FIGURE 6.** The flowchart of GIM. $D_i'$ is the output result of CAM while $U_{i+1}'$ denotes the bilinear interpolation result of $U_{i+1}$. The symbol @ represents the concatenation operation and * represents element-wise addition.

The main function of GIM is to decouple the linear and nonlinear expressions as two branches. The architecture is depicted in Figure 6. The concatenation layers followed by convolution and ReLU activation functions form the non-linear branches. Through this branch, the neural networks can approximate the mapping function well in the training process. The element-wise addition layer is the linear branch of GIM. In a certain sense, this design embeds a “constant” value in the output. Therefore, the existence of the linear layer guarantees the gradient update of each layer to be close to one constant value Compared with the conventional tandem structure to connect linear and nonlinear layers [5], [44], our decoupled module shows better gradient stability during the training process.

Here, we denote the feature maps as $D_i(i = 1, 2, 3, 4, 5)$ and $U_j(j = 1, 2, 3, 4, 5)$ for the down-sampling and up-sampling process respectively. Then GIM can be expressed as follows:

$$U_i = \begin{cases} D_i' + r(\text{concat}(D_i', U_{i+1}')) & i = 1, 2, 3 \\ \text{deconv}(D_5) & i = 4 \end{cases}$$

(4)

where the $D_i'$ denotes the feature map transferred from $D_i$, namely the output of CAM. The $\text{concat}()$ represents feature concatenation and $r$ represents $\text{Conv} + \text{ReLU}$ operations. For the sake of clarity, we expand the formula gradually as follows:

$$M_i = \text{concat}(D_i', U_{i+1}')$$

(5)

$$\tilde{M}_i = \text{ReLU}(\text{conv}(M_i))$$

(6)

$$U_i = D_i' + \tilde{M}_i$$

(7)

where $\text{conv}$ in Eq (6) denotes the $1 \times 1 \times C_i$ convolution operations.

To generate $U_i(i = 1, 2, 3)$, the GIM takes the CAM output $D_i'$ with shape $W_{in} \times H_{in} \times C_{in}$ and the previous up-sampling result $U_{i+1}'$ with the same shape as input. It should be noted that $U_{i+1}'$ is the bilinear interpolation result of $U_{i+1}$, which avoids introducing the nonlinear deconvolution layers.
methodology in PixelLink [11] which predicts the linkage between the pixel and its eight adjacent pixels. For the text/non-text prediction, we utilize softmax function to generate the final feature map with channel $1 \times 2 = 2$, representing text/non-text score respectively; For the pixel-to-pixel linkage prediction, we output the feature map with channel $8 \times 2 = 16$ representing positive/negative scores for every eight adjacent pixels. Then two separate thresholds are applied to generate the positive pixels and the positive linkage. The output segmentation map is obtained by linking the positive pixels using positive linkages. Finally, Connected Components (CC) algorithm is employed to find the individual text areas and minAreaRext in OpenCV [45] is applied to obtain the bounding boxes for every CC.

E. OPTIMIZATION
The overall loss function is given as follows:

$$L = \lambda L_{\text{pixel}} + L_{\text{link}}$$

where $L_{\text{pixel}}$ and $L_{\text{link}}$ denote the pixel loss and the link loss separately. $\lambda$ is set to 2.0 because the pixel classification task is more important than the link prediction task.

Generally, we follow the same methodology in PixelLink [11]. For pixel loss computing, a novel instance-balanced cross-entropy loss is adopted to deal with large difference of the text size, which takes the text instance areas as the modulation factor. Link loss is a kind of class-balanced cross-entropy loss which is calculated on positive pixels only. Specific details can refer to PixelLink [11].

IV. EXPERIMENTS
To evaluate the proposed method, we conduct experiments on three commonly used public datasets: ICDAR2013 (IC13) [46], ICDAR2015 (IC15) [47], MSRA-TD500 (TD500) [48]. The experiment settings and the comparison results are presented as follows. A concise description of these three datasets and the accepted evaluation protocols are given in Section A. The implementation details are illustrated in Section B. Experimental results for ICDAR 2013, ICDAR 2015 and MSRA Td500 are given in Section C, Section D and Section E respectively. To reveal the effectiveness of two proposed modules, sufficient ablation studies are given in Section F.

A. DATASETS AND EVALUATION PROTOCOL
1) SynthText in the Wild (Synth Text) [49]: The SynthText dataset contains 800k synthesized text images which are rendered on natural images. With a proper learning algorithm to choose and transform the text location, the images tend to have a realistic look. Annotations are given in character, word and text line level. This dataset is used to pre-train the model.
2) ICDAR 2015 (IC15) [47]: IC15 is the most commonly used benchmark for detecting multi-oriented text with word-level quadrilaterals. It is composed of 1000 training images and 500 testing images. This dataset is challenging because of the arbitrary orientations, motion blur and low resolutions.
3) ICDAR 2013 (IC13) [46]: IC13 consists of 229 training images and 233 testing images with word-level annotations provided. It is the standard benchmark for evaluating near-horizontal text detection.
4) MSRA TD500 (TD500) [48]: TD500 contains 500 images (300 for training, 200 for testing) containing both Chinese and English texts. Different from IC15 and IC13, annotations of TD500 are at the line level which are represented by the aligned horizontal rectangles.
5) Evaluation protocol: In our experiments, we utilize the standard evaluation protocols for text detection, namely precision (P), recall (R) and f-measure (F). Precisely, they are defined as follows:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F = 2 \times \frac{P \times R}{P + R}$$

where $TP, FP, FN$ denotes the correctly classified text instances, incorrectly spotted instances and the missing instances respectively. Given a detected text instance $b$, it will be assigned to the positive detection if the IOU between $b$ and a ground truth is larger than a given threshold (normally set to 0.5). Since there is a trade-off between recall and precision, f-measure is a more objective measurement for performance assessments.

B. IMPLEMENTATION DETAILS
The experiments are conducted on one single NVIDIA Titan X GPU with 12GB memory. The whole model is optimized by SGD with momentum = 0.9 and weight decay = $5 \times 10^{-4}$. The model is pre-trained on SynthText for 500K iterations which takes about 0.9s per iterations. The learning rate is set to $10^{-3}$ in the pre-trained process. Then we fine-tune the model on IC15, IC13 and TD500 respectively. We set the learning rate to $1 \times 10^{-3}$ for the first 100K iterations and $1 \times 10^{-4}$ for the rest 80K iterations. The input images are resized to $1280 \times 768, 768 \times 768, 512 \times 512$ for IC15, IC13 and Td-500 respectively. We take the data augmentation strategies following SSD and PixelLink. Input images are firstly rotated at a probability of 0.2, by a random angle of $0, \pi/2, \pi$ or $3\pi/2$. Then we randomly crop them with areas ranging from 0.1 from 1, and aspect ratios of the images are ranging from 0.5 to 2. As for the text instance in the processed images, the shorter side less than 10 pixels are ignored.

C. HORIZONTAL TEXT DETECTION PERFORMANCE: IC13
The final models are fine-tuned for about 180K iterations based on the SynthText pre-trained model. Thresholds on the
We have evaluated the proposed GISCA for text detection on ICDAR 2013, a popular horizontal text dataset. The numerical comparison results are listed in Table 1. In general, our proposed STD model outperforms all the existing state-of-the-art methods in terms of recall and f-measure. Specifically, it achieves the best performance compared to all the regression-based methods [13], [14], [16], [18], [20], [22], [31] in three evaluation measurements. Besides, GISCA outperforms all the segmentation based text detection algorithms [11], [50] in terms of recall and f-measure. Taking our baseline model PixelLink as an example, GISCA exceeds PixelLink by 3.9%, 4.2%, 4.0% in recall, precision, f-measure respectively. The only exception is that GISCA has a lower precision value than Mask TextSpotter. Considering that Mask TextSpotter uses the results of text recognition to refine the detection results, it is reasonable to generate such a high precision value. On the whole, the experimental results show the superiority of the proposed GISCA.

### D. MULTI-ORIENTED TEXT DETECTION PERFORMANCE: IC15

We follow the experiment settings presented in section B. The thresholds for pixel and link prediction are both set to 0.8. The results are evaluated by the official submission server for the fair comparison. Some qualitative illustrations are shown in Figure 7(b).

The IC15 text detection task is more challenging compared to IC13 owing to its multi-orientation characteristics in complex contexts. Besides, images in IC15 are of low resolution and contain many small text instances. Table 2 shows results on GISCA compared with previous state-of-the-art published methods. The experiment results reveal the proposed method’s advantages under the complex background and low resolution conditions. More specifically, the proposed method outperforms all the existing algorithms according to the value of recall (85.5%), precision (89.1%) and f-measure (87.3%). Especially in f-measure value, GISCA leads the second-place FTSN by 3.2%. Therefore, GISCA has a huge advantage in detecting multi-oriented text.

### E. ORIENTED MULTI-LINGUAL TEXT LINE DETECTION PERFORMANCE: TD500

MSRA TD500 is the dataset which contains both English and Chinese texts. The annotations are given in the forms of text lines. The thresholds of pixel and link are set to 0.8 and 0.7 respectively.

Some qualitative detection results are illustrated in Figure 7(c) and comparison results are listed in Table 3. We can conclude that GISCA can successfully detect long text lines of arbitrary orientations and shapes. More specifically, GISCA surpasses the existing state-of-the-art results by more than 2% according to the value of f-measure. Although RRD slightly surpasses the proposed method in precision value, the proposed method result is still competitive considering that RRD is a detector specifically designed for long oriented texts.

### F. ABSTRACTION EXPERIMENTS

Our proposed multi-oriented text detection model invokes two innovative components, namely Contextual Attention Module (CAM) and Gradient-inductive Module (GIM). To verify the validity of these two proposed modules, we conduct ablation studies over ICDAR 2013, ICDAR 2015 and MSRA TD500 respectively. Four sets of comparison results are list in Table 4,5,6. For convenience, we utilize Baseline to denote the original PixelLink model which takes the commonly used U-Net as backbone network. The second experiment Baseline+C replaces the skip connection in U-Net with the proposed CAM. The third is Baseline+GIM which follows the architecture of U-Net but it takes the proposed GIM in the up-sampling process. The last is CAM+GIM that incorporates both CAM and GIM in the backbone network.
FIGURE 7. Some qualitative detection results on (a) ICDAR 2013, (b) ICDAR 2015, and (c) MSRA-TD500.

FIGURE 8. The up-sampling process results. The symbol in the upper right corner is the name of the feature map. (a) With CAM. (b) Without CAM. It is obvious that when CAM is integrated into the network, the feature maps filter out the background interference and focus on the text area.


<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>f-measure(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>87.5</td>
<td>88.6</td>
<td>88.1</td>
</tr>
<tr>
<td>Baseline+CAM</td>
<td>91.5</td>
<td>92.4</td>
<td>91.9</td>
</tr>
<tr>
<td>Baseline+GIM</td>
<td>91.4</td>
<td>92.2</td>
<td>91.8</td>
</tr>
<tr>
<td>CAM+GIM</td>
<td>91.4</td>
<td>92.8</td>
<td>92.1</td>
</tr>
</tbody>
</table>

1) WITH OR WITHOUT CAM
As shown in Table 4,5,6, Baseline+CAM outperforms Baseline evaluated by recall, precision and f-measure.

TABLE 5. Ablation experiments of IC15.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>f-measure(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>82.0</td>
<td>85.5</td>
<td>83.7</td>
</tr>
<tr>
<td>Baseline+CAM</td>
<td>84.5</td>
<td>87.9</td>
<td>86.2</td>
</tr>
<tr>
<td>Baseline+GIM</td>
<td>83.5</td>
<td>85.6</td>
<td>84.5</td>
</tr>
<tr>
<td>CAM+GIM</td>
<td>83.5</td>
<td>89.1</td>
<td>87.3</td>
</tr>
</tbody>
</table>

For instance, considering f-measure, Baseline+CAM surpasses Baseline by 3.8%, 2.5%, 2.5% in IC13, IC15 and TD500 respectively. Similarly, CAM+GIM also exceeds
Baseline+GIM on all three measurements. These two sets of comparison experiments show that Contextual Attention Module is an efficient structure in the text detection task. The improvement of the precision value indicates that the CAM effectively reduces the false detection rate and thus alleviates the FP problem. The up-sampling process with and without CAM are shown in the Figure 8, which intuitively displays of the effects of CAM.

2) WITH OR WITHOUT GIM
To validate the potency of GIM, we need to compare Baseline with Baseline+GIM and Baseline+CAM with CAM+GIM. Taking the comparison between Baseline and Baseline+GIM into account, the latter experiments outperform the former ones on all three datasets and under almost all three measurement criteria. The only exception is the Baseline surpasses Baseline+GIM in the value of recall in TD500 which could be attributable to the large text size variation. The other sets of comparison results between Baseline+CAM and CAM+GIM also indicate that GIM is an efficient module in most cases.

V. CONCLUSION
In this work, we have presented an end-to-end trainable scene text detector named GISCA which is under the U-Net framework. More specifically, a novel Contextual Attention Module (CAM) is carefully designed to boost the salient text areas which leads to the reduction of the false positive results. Besides, the proposed Gradient-inductive Module (GIM) is designed to provide a linear bypass which aims at alleviating the gradient vanishing/exploding problems. Experimental results on three widely-used datasets (ICDAR 2013, ICDAR 2015 and MSRA TD500) demonstrate the superiority of the proposed method, which outperforms all state-of-the-art methods in terms of f-measure. In the future, we plan to investigate the universality of CAM and GIM in deep neural networks and probe the curve text detection problems using the proposed method.

ACKNOWLEDGMENT
Special acknowledgements are given to Aalto-PKUSZ Joint Research Center of Artificial Intelligence on Scene Cognition and Technology Innovation for its support.

REFERENCES


TABLE 6. Ablation experiments of TD500.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Recall(%)</th>
<th>Precision(%)</th>
<th>F-measure(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>73.2</td>
<td>83.0</td>
<td>77.8</td>
</tr>
<tr>
<td>Baseline+CAM</td>
<td>75.9</td>
<td>85.3</td>
<td>80.3</td>
</tr>
<tr>
<td>Baseline+GIM</td>
<td>72.5</td>
<td>84.6</td>
<td>78.1</td>
</tr>
<tr>
<td>CAM+GIM</td>
<td>77.1</td>
<td>86.3</td>
<td>81.4</td>
</tr>
</tbody>
</table>


MENG CAO received the B.E. degree from the Huazhong University of Science and Technology (HUST), in 2018. He is currently pursuing the M.Sc. degree in computer science and engineering with Peking University. His current research interests include scene text detection and computer vision.

YUEXIAN ZOU received the M.Sc. degree from the University of Electronic Science and Technology of China, in 1991, and the Ph.D. degree from The University of Hong Kong, in 2000. She is currently a Full Professor with Peking University, where she is also the Director of the Advanced Data and Signal Processing Laboratory, Shenzhen Graduate School. Her current research interests include machine learning for signal processing and deep learning and its applications. She was a recipient of the Leading Figure for Science and Technology Award by the Shenzhen Municipal Government, in 2009.

DONGMING YANG received the B.E. degree from Shanxi University, in 2015, and the M.Sc. degree from the University of Chinese Academy of Sciences, in 2018. He is currently pursuing the Ph.D. degree in computer science and engineering with Peking University. His current research interests include computer vision and pattern recognition.

CHAO LIU received the B.E. degree from the Institute of Disaster Prevention, in 2016. He is currently pursuing the M.Sc. degree with the School of Electronic and Computer Engineering, Peking University. His current research interest includes scene text detection.