Decoupling Recognition from Detection: Single Shot Self-Reliant Scene Text Spotter

Jingjing Wu∗
Harbin Institute of Technology
Shenzhen, China
21s151159@stu.hit.edu.cn

Pengyuan Lyu∗
Department of Computer Vision
Technology, Baidu Inc.
Shenzhen, China
lvpyuan@gmail.com

Guangming Lu
Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
Harbin Institute of Technology
Shenzhen, China
luguangm@hit.edu.cn

Chengquan Zhang
Department of Computer Vision
Technology, Baidu Inc.
Shenzhen, China
zhangchengquan@baidu.com

Kun Yao
Department of Computer Vision
Technology, Baidu Inc.
Shenzhen, China
yaokun01@baidu.com

Wenjie Pei†
Harbin Institute of Technology
Shenzhen, China
wenjiecoder@outlook.com

ABSTRACT
Typical text spotters follow the two-stage spotting strategy: detect the precise boundary for a text instance first and then perform text recognition within the located text region. While such strategy has achieved substantial progress, there are two underlying limitations. 1) The performance of text recognition depends heavily on the precision of text detection, resulting in the potential error propagation from detection to recognition. 2) The RoI cropping which bridges the detection and recognition brings noise from background and leads to information loss when pooling or interpolating from feature maps. In this work we propose the single shot Self-Reliant Scene Text Spotter (SRSTS), which circumvents these limitations by decoupling recognition from detection. Specifically, we conduct text detection and recognition in parallel and bridge them by the shared positive anchor point. Consequently, our method is able to recognize the text instances correctly even though the precise text boundaries are challenging to detect. Additionally, our method reduces the annotation cost for text detection substantially. Extensive experiments on regular-shaped benchmark and arbitrary-shaped benchmark demonstrate that our SRSTS compares favorably to previous state-of-the-art spotters in terms of both accuracy and efficiency.

CCS CONCEPTS
• Applied computing → Optical character recognition; • Computing methodologies → Scene text spotting.

KEYWORDS
OCR (optical character recognition), text detection, text recognition

ACM Reference Format:
Jingjing Wu, Pengyuan Lyu, Guangming Lu, Chengquan Zhang, Kun Yao, and Wenjie Pei. 2022. Decoupling Recognition from Detection: Single Shot Self-Reliant Scene Text Spotter. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548266

1 INTRODUCTION

Figure 1: Comparison between typical two-stage methods (a) and our proposed SRSTS (b). Instead of recognizing text from the detected and cropped RoIs as previous two-stage methods, the detection module and recognition module in our SRSTS are decoupled and bridged by positive anchor points predicted by anchor estimation module. As a result, our recognition module can recognize texts correctly even though the detection is not precise. ‘s’ in green are the estimated anchor points and blue points denote the sampled points.

Scene text spotting aims to recognize all the texts appearing in an image, which is a highly challenging task due to diverse text appearances in terms of shape, size and style. Most of existing methods for text spotting [8, 15, 16, 21–23, 28, 34] follow the two-stage spotting strategy, which performs text detection to locate the precise boundaries of text instances in the first stage and then recognizes the text within the detected RoI in the second stage. While remarkable progress has been made by these methods, there are two potential limitations preventing those methods from achieving better performance. One is that the performance of text recognition relies
heavily on the precision of text detection, as a result, such methods suffer from the error propagation from text detection. The other is that the RoI cropping operation may carry some noises from background and bring information loss caused by the operation of pooling or interpolation from feature maps, which also degrade the performance of recognition module. Consider the example in Figure 1, the curved text instance ‘ristorante’ is extremely difficult for detection and the irregular shape is also not so friendly to RoI cropping. Accordingly, the previous two-stage text spotters fail to locate the accurate boundary during text detection. Besides, the RoI cropping operation may also bring some errors when conducting feature sampling. Consequently, the characters ‘r’ and ‘e’ are missed or wrongly recognized in the decoding stage.

Recently, some single shot text spotters [26, 35, 38] which integrate both the text detection and recognition into one-stage are proposed. Benefiting from the one-stage architecture, the RoI cropping operation is not necessary and the disadvantages of RoI cropping are avoided. However, these methods perform recognition depending on accurate detection of text region [38] or center line of text instances [35], which still have the limitation of error propagation from detection to recognition. In [26], Qiao et al. proposed MANGO which treats text spotting task as a pure text recognition problem. However, extra character-level annotation for supervision is required by [26], which is extremely label-consuming.

In this work we propose to decouple recognition from detection and thereby reduce the interdependence between them. To be specific, we propose the single shot Self-Reliant Scene Text Spotter (SRSTS). Inspired by the concept of ‘anchor’ in [29], we introduce anchor point that denotes each pixel of feature maps which serves as reference point for detection and recognition task with location guidance. SRSTS estimates positive anchor points for each potential text instance and the obtained positive anchor points are further used to bridge the detected candidates and the recognized texts. As for the text recognition, we perform sampling for each anchor point and further decode text based on the features of the sampling points. As a result, the precise text detection is not essential for text recognition, leading to two prominent merits. First, the performance of recognition is not strictly limited by the precision of detection, which is particularly advantageous in challenging scenarios for detection since estimating a rough anchor point for a text instance is a lot easier than predicting its precise boundary. Secondly, since our recognition task does not relay on precise detection results, our SRSTS reduces the annotation cost for text detection substantially than the typical two-stage spotting methods, let alone the character-level annotations required by the previous single shot spotters [26, 38]. Figure 1 shows that our SRSTS is able to recognize the curved text correctly based on the sampled points around the estimated anchor point, even though our method also cannot detect its boundary precisely. To conclude, we make the following contributions.

- Our proposed SRSTS reduces the dependence of text recognition on detection, thereby circumventing the potential error propagation from detection to recognition.
- Benefiting from the designed effective spotting mechanism, SRSTS is able to recognize the text instances correctly even though the text boundaries are challenging to detect.
- We conduct extensive experiments to evaluate our method, including various ablation studies to investigate the effectiveness of our method and the comprehensive comparisons to the state-of-the-art methods for text spotting on both regular-shaped benchmark and arbitrary-shaped benchmark. All these experiments validate the advantages of our method.

2 RELATED WORK

The previous scene text spotters can be coarsely divided into two categories: two-stage scene text spotters and single shot scene text spotting methods.

Two-stage scene text spotting. In the early stage, the scene text spotters [12, 17, 18] are always composed of separate text detector and recognizer. Jaderberg et al. [12] use a regression-based text detector and a word classification based recognizer to detect text and recognize text respectively. In [17, 18], a SSD [19] based text detector and CRNN [30] are employed to read text. Though substantial progress has been achieved, those methods still struggle to get better performance due to the sub-optimal between text detection and recognition.

To mitigate the sub-optimal between text detection and recognition, some end-to-end trainable frameworks are proposed. Busta et al. [4] first train text detector and recognizer separately and then fine-tune the two module jointly. In [14], to learn text detector and recognizer in an end-to-end manner, a complex curriculum learning paradigm is used. To train text detection and recognition jointly and steadily, some methods [20–22, 28, 36] use the groundtruth boxes to extract feature for text recognition. In [15, 16, 23], the above mentioned issue is relieved in the way of character segmentation.

Detecting and recognizing text of arbitrarily shapes has also attracted a lot of attention. Mask TextSpotter [23] reads text by segmenting text regions and characters. Qin et al. [28] proposed RoI masking which multiplies the cropped features with text instance segmentation masks. TextDragon [8] uses RoISlide to sample local features along the centerline. Mask Textspotter v3 [16] proposed SPN (segmentation proposal network) to represent arbitrary-shape proposals. Text Perceptron [27] and Boundary [34] represent the text boundary as a group of key points, and apply the TPS (thin-platespline Transformation) [3] to rectify irregular boundary. ABCNet [21] and ABCNet v2 [22] innovatively use Bezier curve to fit curve contour and introduce BezierAlign operation to crop text feature map.

Most of the above mentioned two-stage methods conduct text detection and recognition serially and connect them by the RoI cropping operation. As a result, the recognition result is heavily affected by detection and RoI cropping.

Single shot scene text spotting. Recently several works attempt to integrate the detector and recognizer into a one-stage network to avoid the adverse effects of RoI cropping. CharNet [38] simultaneously predicts instance-level position information and character-level bounding box with character label. Based on instance-level detection result, the text instance can be generated by grouping predicted characters. Recently PGNNet [35] predicts various text instance information in parallel and adopts point-gathering operation to gather the pixel-level character classification probability. To further reduce the dependence on detection module, MANGO [26] treats
Decoupling Recognition from Detection: Single Shot Self-Reliant Scene Text Spotter

Figure 2: Architecture of the proposed single shot Self-Reliant Scene Text Spotter. It consists of four modules: Feature Extractor, Anchor Estimator, Text Detector and Text Recognizer. Taking extracted multi-scale features from Feature Extractor, Anchor Estimator predicts positive anchor points for each potential text instance to perceive its rough position. Text Detector locates precise text boundary and Text Recognizer performs sampling referring to each anchor point and further decodes text based on the features of the sampled points. The detected polygons and decoded texts are further bridged by positive anchor points.

the spotting task as a pure recognition task. However, MANGO requires character-level annotations, and can’t be utilized for real-time application.

Our method is also a single shot scene text spotter. We bridge the detection task and recognition task by anchor points and predict detection and recognition results in parallel, resulting in the decoupling between detection and recognition. Our method outperforms the previous one-stage methods in the three major advantages: 1) compared to [35], our method decouples recognition from detection, thus the recognition does not rely on precise detection results; 2) compared with [26, 38], ours method only needs word-level annotations, while [26, 38] require character-level annotations; 3) our method is able to run at real-time, while [26, 38] can not.

3 SELF-RELIANT SCENE TEXT SPOTTER

Unlike typical scene text spotters that rely on text detection to acquire the precise boundaries of text instances for text recognition, the proposed Self-Reliant Scene Text Spotter (SRSTS) takes anchor points as reference and conducts text detection and recognition in parallel. As a result, our SRSTS is able to decouple text recognition from text detection and reduce the dependence of text recognition on the detection performance, thereby circumventing the potential error propagation from detection to recognition. We will first present the overall framework of the SRSTS, and then describe how to estimate positive anchor points for text instances. Finally we will elaborate on the text detection and text recognition.

3.1 Overall Framework

As illustrated in Figure 2, our SRSTS consists of four modules: Feature Extractor, Anchor Estimator, Text Detector and Text Recognizer. Given a text image, the SRSTS first employs the Feature Extractor to learn multi-scale features, and then conducts positive anchor points estimation, text detection and text recognition at the same time. The Anchor Estimator estimates positive anchor points for potential text instances. In the inference stage, for each obtained positive anchor point indicating a text instance, the Text Detector locates precise text boundary, and the Text Recognizer of SRSTS performs sampling around positive anchor point and then performs text decoding based on the features of the sampling points.

Feature Extractor. The Feature Extractor of the SRSTS adopts the similar network structure to BiFPN [32], which is composed of a downsampling pathway and an upsampling pathway. The downsampling pathway employs a ResNet-50 [11] as the feature learning backbone while the upsampling pathway generates multi-scale feature maps by fusing the features from both pathways in the corresponding feature levels. To be specific, we use two different scales of features (P₂ and P₃ in Figure 2) in the downstream for Anchor Estimator, Text Detector and Text Recognizer, whose feature map size are \( \frac{1}{4} \times \frac{1}{4} \) and \( \frac{1}{8} \times \frac{1}{8} \) of the input image size, respectively. Larger feature map has smaller receptive field is used for spotting smaller size of text instances.

Anchor Estimator. Instead of spotting text serially as the previous methods, our SRSTS decouples recognition from detection by conducting text detection and recognition separately with the reference of anchor points. The anchor points play an important role in that: 1) Text detection and text recognition are conducted with the location guidance of anchor points. 2) a positive anchor point serves as a reference point to indicate the existence of a text instance around it. 3) a detected polygon can be matched with a recognized text by the shared positive anchor point. We estimate positive anchor points with Anchor Estimator.
Text Detector. Based on the anchor points, text detection is performed with the Text Detector. For the consideration of generalization, efficiency and accuracy, we adopt a real-time instance segmentation framework YOLACT [2] to detect text of arbitrary shapes.

Text Recognizer. The anchor points are further used to guide the Text Recognizer of the SRSTS to decode text. The Text Recognizer performs sampling for each anchor point by a designed sampling module and then decodes the text of this instance based on the features of the sampling points.

3.2 Estimation of Positive Anchor Points

The Anchor Estimator is designed to predict positive anchor points for each potential text instance, which serve as the reference positions for text detection and text recognition in the post-processing. A confidence map is learned to quantify the probability for each anchor point of feature maps.

Learning confidence map for anchor points. The confidence map is learned from the feature maps output from the Feature Extractor. Each pixel value in the confidence map, which is within [0, 1], indicates the probability for this pixel to be a positive anchor point.

Formally, given a feature map $P_i \in \mathbb{R}^{W \times H \times C}$ containing $C$ channels of features with size $W \times H$, a non-linear transformation is performed by a $3 \times 3$ convolutional layer and a $1 \times 1$ convolutional layer with Batch Normalization and ReLU in between. Here $1 \times 1$ convolutional layer is applied to reduce the channel number to 1. Finally, Sigmoid function $\sigma$ is used to project confidence values into $[0, 1]$. Mathematically, the confidence map $C_i \in \mathbb{R}^{W \times H}$ is calculated from the feature map $P_i$ by:

$$C_i = \sigma\left(\mathcal{F}_{\text{conv}1 \times 1}\left\{\text{BN}\_\text{ReLU}(\mathcal{F}_{\text{conv}3 \times 3}(P_i))\right\}\right)$$

where $\mathcal{F}_{\text{conv}1 \times 1}$ and $\mathcal{F}_{\text{conv}3 \times 3}$ denote the operations of $1 \times 1$ and $3 \times 3$ convolutional layers, respectively.

Intuitively, a positive anchor point should be close to the center position of its corresponding text instance. Thus, we optimize the parameters for learning the confidence map (in Equation 1) in a supervised way with the Dice loss [24]:

$$\mathcal{L}_{\text{c}} = 1 - \frac{2|C \cap C_{\text{gt}}|}{|C| + |C_{\text{gt}}|}.$$  \hspace{1cm} (2)

Herein, $C_{\text{gt}}$ is the groundtruth for the confidence map, which is derived from the detection groundtruth of text instances represented by bounding boxes or polygons. Specifically, the pixels of the central region (with tuned area) for each text instance in $C_{\text{gt}}$ are assigned the value of 1 while other pixels are assigned 0. The Dice loss is used to maximize the overlap between the predicted confidence map $C$ and the groundtruth $C_{\text{gt}}$.

Multi-scale confidence maps for capturing multi-scale text instances. Each of multi-scale feature maps from the Feature Extractor produces an individual confidence map and each confidence map has a corresponding groundtruth. We assign text instances to groundtruth maps heuristically, and each text instance (in the input image) only appears once in all groundtruth confidence maps. Specifically, larger text instances are assigned to the groundtruth confidence map with smaller size since the corresponding feature map has larger receptive field and is favorable for spotting larger size of text instances. As a result, we learn multi-scale confidence maps, each of which being responsible for the appropriate size of text instances. The size thresholds for the assignment of text instances to groundtruth confidence maps are hyper-parameters tuned on a validation set.

Label generation. Since the height of the text generally determines the size of the characters, we assign text instances to feature maps at different levels according to the height of text instances. More specifically, we compute the average height $h^*$ for each individual text instance at first. If $h^* > m_i$ and $h^* < m_{i+1}$, it is regarded as a positive sample in level-$i$ and a negative sample in the other level. In this work, the $m_2, m_3$ and $m_4$ are set to 0, 40 and $\infty$. When a text instance is assigned to level-$i$, its positive anchor points only appear in the corresponding ground-truth of level-$i$.

We generate the groundtruth map in a similar way to [40]. We define the positive region of the confidence map as the central region of text instance. Concretely, for each text instance, we firstly find the center axis, then roughly shrink the quadrilateral whose corners are the two pairs of points close to the center axis. For regular text instance, the shrink ratio is 0.2. As the quadrilateral is shorter than the original text polygon for most curved text instance, the height of the quadrilateral is shrunk to 0.2 while the shrink ratio of width is enlarged but no more than 1. The shrink ratio of width is adjusted according to the following formula:

$$\text{ratio} = \min(1, 0.1 \times (N_p/2 - 1))$$\hspace{1cm} (3)

$N_p$ denotes the number of polygon vertexes. For curved text instance, $N_p \geq 6$. The shrink version of quadrilateral is regarded as the positive region in which all pixels are labeled 1.

3.3 Text Detection Guided by Anchor Points

With the location guidance of anchor points, we conduct text detection to obtain the boundary of text with arbitrary shape. Following YOLACT [2] which shares the predicted mask, our text detector can detect text with arbitrary shapes flexibly and efficiently.

As shown in Figure 2, bounding box regressor, mask coefficient predictor and protonet are used to generate bounding box, mask coefficients and prototypes in parallel. Specifically, fed a feature map $P_i \in \mathbb{R}^{W \times H \times C}$ to bounding box regressor and mask coefficient predictor which consist of similar structure as anchor estimator, geometry map $G_i \in \mathbb{R}^{W \times H \times 4}$ and mask coefficient map $M_i \in \mathbb{R}^{W \times H \times k}$ are yielded. Here, the 4 channels in geometry map represent 4 distances from the anchor point to the top, right, bottom and left boundaries of the rectangle of a text instance respectively. And $k$ in mask coefficient map means the number of mask coefficients, which is set to 4 by default. For the prototypes generation, we input the $P_i$ to protonet which contains stacked convolutional layers as anchor estimator and resize the output prototypes to the same shape as the input image. As a result, $k$ shared prototypes $T_i \in \mathbb{R}^{W \times H \times k}$ as well as $k$ mask coefficients for each anchor point are yielded, where $W$ and $H$ are the shape of the input image. In the end, the final text mask $F \in \mathbb{R}^{W \times H}$ for each instance is generated by linearly combining prototypes and mask coefficients as [2].

Similar to [37, 40], we apply IOU Loss [39] as $\mathcal{L}_{\text{iou}}$ to train bounding box regressor and adopt Dice Loss for $\mathcal{L}_{\text{mask}}$ to optimize
protonet and mask coefficient predictor. In the inference stage, we perform Non-maximum Suppression (NMS) based on the bounding box and confidence value of positive anchor point to remove the redundant candidates, and only one positive anchor point is kept for each potential text instance. And then the text masks of kept candidates are binarized, and the predicted text polygons are obtained from the connected regions of the binary maps.

3.4 Text Recognition Guided by Anchor Points

The anchor points are further used to guide the Text Recognizer of our SRSTS to decode text. Without relying on the detected bounding boxes or polygons for defining exact regions for a text instance, the Text Recognizer performs sampling around anchor points. Then the Text Recognizer decodes texts based on the features of the sampling points.

Weakly supervised sampling around anchor points. As illustrated in Figure 2, given an anchor point, the Text Recognizer performs sampling by predicting the two-dimensional coordinates of each sampled point with a designed sampling module. The sampling module consists of three $3 \times 3$ convolutional layers and one $1 \times 1$ convolutional layer, with the Batch Normalization and ReLU function between the layers. Taking a feature map $P_i \in \mathbb{R}^{W \times H \times C}$ from the Feature Extractor as input, the sampling module samples $K$ points for each anchor point in $P_i$ and outputs a tensor $S_i \in \mathbb{R}^{W \times H \times 2K}$:

$$S_i = f_s(P_i),$$

where $f_s$ denotes the transformation function of the sampling module. Thus, the coordinates of the $K$ sampled points for an anchor point at $[w, h]$ in level-1 are specified by the vector $S_i[w, h]$. As the sequence lengths in English words are almost not very long, we set $K$ to 25 empirically as [30].

Intuitively, high-quality sampled points around an anchor point are expected to involve all characters in the text instance indicated by this anchor point. To this end, we conduct supervision on the sampling module to encourage it to sample uniformly along the center line of the text instance:

$$L_s = ||S - S_0||_1,$$

where $S_0$ denotes the groundtruth, which is the coordinates of uniformly distributed $K$ points along the center line of the text instance. $S_0$ can be easily calculated based on the groundtruth polygon of this text instance.

Note that sampling along the center line of the polygon is just one of optional ways, rather than the sole way, to improve the sampling quality. Thus we only perform such supervision during the pre-training stage to guide the sampling process and speed up the convergence, which is equivalent to a weak supervision. Additionally, the sampling process is also supervised indirectly by the text decoding loss in Equation 7 to learn proper sampling distributions.

Text decoding. For each anchor point, $K$ points are sampled around it to capture the key features for recognizing all characters of this text instance. Then the Text Recognizer decodes the text based on the features of these sampled points. Specifically, the Text Recognizer performs character classification among 37 character classes (36 for alphanumeric characters and 1 for the blank) for each sampled point taken from the Feature Extractor, using a character classification head $f_{cls}$. For an anchor point, the classified probability distribution $p_k$ for the $k$-th sample is obtained by:

$$p_k = f_{cls}(P_i),$$

where $P_i$ is the feature map where this anchor point appears in. The character classification head $f_{cls}$ is constructed with four $3 \times 3$ convolutional layers and $1 \times 1$ convolutional layer in the end. ReLU function is used as activation function between layers. In practice, the character classification head predicts the probability distribution for all pixels in a feature map in parallel for efficiency whilst only the probability of sampled points is used during text decoding. Gathering all predicted probability distributions for all sampled points w.r.t. each of anchor points, the Text Recognizer decodes the text under the supervision of CTC loss [9]:

$$L_{ctc} = \sum_{n=1}^{N} \text{CTC}_\text{loss}(p_{n \alpha}, q_{n \alpha}).$$

Herein, $p_{n \alpha}$ is the gathered probability distribution sequence of all $K$ samples for the anchor point $n$ among total $N$ positive anchor points. $q_{n \alpha}$ is the corresponding groundtruth. Such way of decoding text based on the probabilities of sequential sample points has been explored in PGNet [35].

Rationale. Typical two-stage text spotters depend on text detection to locate the precise text boundaries for text recognition, which implies that the performance of text recognition heavily relies on the detection performance. We argue that the accurate detection of text boundaries is not necessary for text recognition. In contrast to such traditional way, our SRSTS samples points around anchor points to capture the key features for text decoding. As a result, the precise text detection is not essential for our method, which yields two advantages for our method: 1) less dependence between text recognition and detection and thus less error propagation from detection to recognition; 2) lower annotation cost for text detection.

3.5 End-to-End Parameter Learning

Following the routine training paradigm adopted by most existing text spotters [22, 26, 36], we first pre-train our SRSTS on two synthetic datasets (‘Synthtext’ [10] and ‘Bezier Curve’ [21]), then fine-tune the model on a mixture of synthetic and real-world datasets. The whole model is optimized jointly in an end-to-end manner:

$$L = \lambda_1 L_c + \lambda_2 L_{iou} + \lambda_3 L_{mask} + \lambda_4 L_s + \lambda_5 L_{ctc},$$

where $L_{iou}$ [39] and $L_{mask}$ are detection losses. $\lambda_1$ to $\lambda_5$ are hyper-parameters to balance between different losses. Note that $\lambda_3$ is equal to 0 in the fine-tuning stage since the sampling supervision $L_s$ is only used in the pre-training stage. Besides, the $L_{iou}$, $L_{mask}$, $L_s$, $L_{ctc}$ only work on the positive anchor points. To balance the different tasks, we set $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ and $\lambda_5$ to 5, 5, 5, 1 and 1 respectively by default.

4 EXPERIMENTS

4.1 Experimental Setup

Benchmarks. We test our model on two challenging benchmarks: ICDAR 2015 [13] and Total-Text [6]. ICDAR 2015 contains 1000 training images and 500 testing images. It is annotated with quadrangles and work-level text transcriptions. For text spotting task, ICDAR 2015 provides 3 lexicons named ‘Strong’, ‘Weak’ and ‘Generic’
Table 1: Comparison between different supervision modes for learning the sampling module. 'None' represents the performance without using lexicon while 'Full' implies using the lexicon containing all words appearing in the test set.

| Supervision mode     | None  | Full |
|----------------------|-------|------|
| Unsupervised sampling| 76.35 | 85.24|
| Weakly supervised sampling | 78.80 | 86.33|
| Fully supervised sampling | 76.49 | 84.59|

Datasets for jointly training. Following the previous methods [22, 26], we combine the following datasets for jointly training: Synthtext is a synthetic dataset provided by [10], which contains 800k images. Bezir Curve Synthetic Dataset 150k is a curved synthetic dataset provided by [21], which contains 90k synthetic straight text images and 50k curved text images. COCO-Text [33] is a real word dataset which contains 63686 images. ICDAR 2017 MLT [25] is a multi-language scene text dataset and only the English samples are used in our training set. ICDAR 2019 ArT [7] contains 5,603 training images. The joint training dataset is composed of the above training datasets and the training set of benchmarks.

Implementation details. The training process is divided into pre-training stage and fine-tuning stage. The model is firstly trained with synthetic datasets and then fine-tuned with joint training set. In particular, we first pre-train our model on Synthtext and Bezir Curve Synthetic Dataset for 4 epochs and then fine-tune the model another 500000 steps with the joint training dataset. The joint training dataset is a mixture of Synthtext, Bezir Curve Synthetic Dataset, COCO-Text, ICDAR 2017 MLT(only English samples) , ICDAR 2019 ArT, ICDAR 2015 and Total-Text. For the fine-tuning task on ICDAR2015 and Total-Text, we set the sample ratio of different datasets to 1:0:1:1:0:1:0 and 1:1:1:2:1:2, respectively.

We use data augmentation to train our model. Specifically, we resize the input image with a randomly selected scale from 0.4 to 1.7 and keep the aspect ratio unchanged. To handle rotated text well, we also randomly rotate the input image with an angle in the range of [-10, 10]. We randomly crop patches from the input image and resize the longer side to the size of 640 and pad the resized image to 640 × 640 for effective training. In addition, some random blur and color jitter are also used.

We use SGD optimizer to optimize our model with the initial learning rate of 2e-3 and 1e-3 for the pre-training and fine-tuning stage, respectively. We set the weight decay to 0.0001 and momentum to 0.9, and delay the learning rate with a ‘poly’ learning rate strategy as [5]. Our model is trained with a batch size of 16, and evaluated with a batch size of 1.

4.2 Ablation Study

In this section, we conduct ablation studies on Total-Text dataset to verify the effectiveness of key components of proposed method.

The effectiveness of the proposed weak supervision of sampling module. We conduct experiments to verify the effectiveness of the proposed weak supervision of sampling module. We train our model on Total-Text dataset with different supervision modes: unsupervised sampling, weakly supervised sampling and fully supervised sampling. Unsupervised sampling means the sampling module is trained only with supervision from recognition loss. Weakly supervised sampling means we only perform direct supervision \( L_s \) during the pre-training stage and in the fine-tuning stage, the sampling module is optimized in a weakly supervised way. Fully supervised sampling denotes the supervision of sampled center lines is used both in pre-training and fine-tuning stage.

The results are listed in Table 1. As shown, even with unsupervised sampling, our model still achieves excellent results. Specifically, without help from lexicon, the F-Measure of our model is 76.35%. When using the 'Full' lexicon, our model achieves the F-Measure of 85.24%. Benefiting from the weakly supervised sampling, the performance of our model can be further improved and 2.45% and 1.09% improvements are obtained. However, with fully supervision, the performance of proposed method decreases slightly. When evaluated without lexicon, the F-Measure of fully supervised sampling is 2.31% worse than weakly supervised sampling.

These results demonstrate the effectiveness of weakly supervision strategy of sampling module. We argue that by adding sampling supervision in the pre-train stage, the proposed SRSTS can be better guided by a direct supervision. And in the fine-tuning stage, we optimize the sampling module by data driven, and the sampling module samples better sample points adaptively.

Training without box or polygon annotations. The recognition module of our model is guided by anchor points and decoupled from the detection module, thus brings two advantages in terms of training data: 1) our model can be trained without the precise annotations of detection. 2) it’s cheaper to obtain an anchor point level annotation rather than precious boundary information for single text instance, and better performance can be achieved by our model when trained with those weakly labeled data.

We conduct a comparative experiment to verify that by adding anchor point labeled images, the spotting performance of proposed method can be significantly improved. We randomly split Total-Text training set into subset 'A' and subset 'B' evenly. While the samples in subset 'A' are labeled with polygons and word-level transcriptions, each individual in subset 'B' only provides a center point and word-level transcription. We train two model with the same settings except the training dataset. The results are shown in Table 2. The 'Base' set is composed of subset 'A' and the left data in joint training set. By adding subset 'B', the E2E performance increases 1.68% in terms of F-Measure and meanwhile the detection result is slightly improved. The results demonstrate that our
Table 3: Quantitative results on the ICDAR 2015 dataset. Methods marked with ‘∗’ are trained with character-level annotations. ‘S’, ‘W’ and ‘G’ denote three levels of informative lexicons (from high to low) used for inference. ‘S’ (strong) means a customized lexicon of 100 words, including the groundtruth, are given for each image. ‘W’ (weak) implies a lexicon includes all words that appear in the test set provided. ‘G’ (generic) denotes a generic lexicon with 90k words.

| Lexicon     | Method               | Detection | E2E       | Word Spotting |
|-------------|----------------------|-----------|-----------|---------------|
|             |                      | R | P | F | S | G | W | S | W | G |
| Official Lexicons | TextNet [31] | 85.41 | 89.42 | 87.37 | 78.66 | 74.90 | 60.40 | 82.38 | 78.43 | 62.36 |
|             | FOTS [20]           | 85.17 | 91.00 | 89.79 | 81.09 | 75.90 | 60.80 | 84.68 | 79.32 | 63.29 |
|             | TextDragon [8]      | 83.75 | 92.45 | 87.88 | 82.54 | 78.34 | 65.15 | 86.22 | 81.62 | 68.03 |
|             | Qin et al. [28]     | 85.75 | 89.36 | 87.52 | 83.38 | 79.94 | 67.98 | -     | -     | -     |
|             | TextPerceptron [27] | 82.50 | 92.30 | 87.10 | 80.50 | 76.60 | 65.10 | 84.10 | 79.40 | 67.90 |
|             | PGNet [35]          | 84.80 | 91.80 | 88.20 | 83.30 | 78.30 | 65.30 | -     | -     | -     |
|             | Boundary [34]       | 87.50 | 89.80 | 88.60 | 79.70 | 75.20 | 64.10 | -     | -     | -     |
|             | PAN++ [36]          | 83.90 | 91.40 | 87.50 | 82.70 | 78.20 | 69.20 | -     | -     | -     |
|             | SRSTS (Ours)        | 81.96 | 96.05 | 88.44 | 82.83 | 80.70 | 69.47 | 87.37 | 84.75 | 72.33 |
| Specific lexicons | CharNet* [38]      | 88.30 | 91.15 | 89.70 | 80.14 | 74.45 | 62.18 | -     | -     | -     |
|             | CRAFTS* [1]         | 85.30 | 89.00 | 87.10 | 83.10 | 82.10 | 74.90 | -     | -     | -     |
|             | MANGO* [26]         | -    | -     | -     | 81.80 | 78.90 | 67.30 | 86.40 | 83.10 | 70.30 |
|             | MaskTextspotter v2 [15] | 87.30 | 86.60 | 87.00 | 83.00 | 77.70 | 73.50 | 82.40 | 78.10 | 73.60 |
|             | MaskTextSpotter v3 [16] | -    | -     | -     | 83.30 | 78.10 | 74.20 | 83.10 | 79.10 | 75.10 |
|             | MANGO* [26]         | -    | -     | -     | 85.40 | 80.10 | 73.90 | 85.20 | 81.10 | 74.60 |
|             | ABCNet v2* [22]     | 86.00 | 90.40 | 88.10 | 82.70 | 78.50 | 73.00 | -     | -     | -     |
|             | SRSTS (Ours)        | 81.96 | 96.05 | 88.44 | 85.63 | 81.74 | 74.51 | 85.84 | 82.61 | 76.82 |

proposed SRSTS reduces the annotation cost for text detection substantially.

4.3 Comparison with State-of-the-art Methods

4.3.1 Results on ICDAR 2015. We conduct experiment on ICDAR 2015 to verify the effectiveness of SRSTS on oriented text. The detailed results are shown in Table 3. In the inference stage, the longer side of input image is resized to 1920 while keeping the aspect ratio unchanged. As character-level annotations are hard to obtain and can rise the performance markedly, we divide existing text spotters into two groups for fair comparison. Methods in the first group only need word-level annotations and the remain have to be trained with character-level annotations. In addition, private lexicons are proposed in [15] and followed by some other methods [16, 22, 26]. To compare comprehensively with others, we also report our results with the specific lexicons.

As shown in Table 3, our method achieves comparable detection results and the best performance on word spotting and end-to-end tasks when compared with the methods only trained with word-level annotations. Besides, though only trained with word-level annotations, our method also outperforms some character-level annotations based methods. In particular, our method surpasses CharNet [38] and MANGO [26] by 7.29% and 2.17% in terms of F-Measure when evaluated with generic lexicon. We also compare our methods with [15, 16, 22, 26] which use specific lexicons, the dominant results of all evaluation situations also show the effectiveness of our method.

4.3.2 Results on Total-Text. To evaluate the effectiveness of SRSTS on arbitrary-shaped text, we conduct experiment on Total-Text. The detailed results are shown in Table 4.

Our method achieves the best end-to-end performance with or without lexicon. Particularly, our method achieves the F-measure of 78.80% and 86.33% respectively, outperforms all the other competitors. Compared with the previous best word-level annotations based method [16], our method is beyond it by a large margin by 7.6% and 7.93% in terms of F-measure. Besides, our methods also surpasses the previous state-of-the-art [1] by 0.1%, though character-level annotations are used and impressive results have been achieved by [1].

4.3.3 Comparison of Speed. Since the evaluation environments of different methods are various and the ways of calculating the speed are also diverse, a fair comparison is necessary. We retest the speed of some real-time text spotters [21, 35, 36] in the same hardware (3090 Ti GPU) and calculate the time cost from inputting the input image to outputting the final result. As shown in Table 4, when the input size is set to 512, our method can run at 22.9 FPS which is 1.49x faster and 13.8% better than the previous fastest method [35]. With larger input size (shorter size of input image is 640), SRSTS can achieve the state-of-the-art performance and still runs faster than all the previous real-time text spotters.

4.4 Qualitative Evaluation

We visualize the detection results, sampling points and spotting results of multi-oriented and curved texts in Figure 3. As shown, the detection fails in challenging scenarios, leading to imprecise boundaries and incomplete detected regions. However, benefiting from the decoupling between detection and recognition, our method can still predict the correct text by sampling the key features for recognition.

5 CONCLUSION

In this paper we have presented the single shot Self-Reliant Scene Text Spotter (SRSTS), which decouples recognition from detection to circumvent the error propagation from detection to recognition. To be specific, SRSTS conducts text detection and recognition in
parallel and bridges the detected polygons and decoded texts by the shared positive anchor points. As a result, the precise detection of text boundaries, which is demanded by typical two-stage text spotters, is not essential for our SRSTS. This merit enables our method to recognize the text correctly even though the text boundaries are challenging to detect. The extensive experiments demonstrate the effectiveness and advantages of our proposed method.

ACKNOWLEDGMENTS
This work was supported in part by the NSFC fund (U2013210, 62006060, 62176077), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant (2019B1515120055, 2022A1515010306), in part by the Shenzhen Key Technical Project under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant (JCYJ20210324132210025), in part by the Shenzhen Stable Support Plan Fund for Universities (GKWD20201230155427003-20200824125730001), in part by CAAI-Huawei MindSpore Open Fund(CAAIXSJLJ-2021-003B), in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China, and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).
A THE GENERALITY OF SAMPLING MODULE

Figure 4: Qualitative results of CTC-based decoder and attention-based decoder, where (a) are the sampled points and spotting results of CTC-based decoder and (b) are that of attention-based decoder. The blue text refers correct sequence output while the red text refers incorrect sequence output.

Table 5: Comparison between CTC-based recognition module and attention-based recognition module on Total-Text.

| Recognition Module | Detection | E2E |
|--------------------|-----------|-----|
|                    | R        | P   | F   | None | Full | FPS |
| CTC                | 82.96    | 91.99 | 87.24 | 78.80 | 86.33 | 18.74 |
| Attention          | 80.71    | 94.60 | 87.10 | 76.01 | 84.81 | 11.74 |

We study the generality of our proposed sampling module and prove that our sampling module can work well with both CTC-based decoder and attention-based decoder. In detail, we replace the CTC-based decoder with attention-based text recognizer as [22], and the comparison is shown in Table 5.

When training with attention-based recognition module, our method achieves comparable detection results as CTC-based model. Besides, the attention-based recognition model achieves close end-to-end recognition results which are 2.79% and 1.52% worse than the CTC-based model. However, the CTC-based model has a significant advantage over the attention-based model in that the CTC-based model can run at 18.74 FPS which is about 1.6x faster than the attention-based model.

To further analyze the reasons for the performance difference between different decoders, we visualize the spotting results and sampled points. The spotting results on curved texts are shown in Figure 4. As can be easily seen, the distribution of sampling points of the attention-based decoder is a bit worse than that of the CTC-based decoder while the detection results are similar. With attention-based decoder, the sampled points fail to cover the curved word instance and thus, 'MIRVISH' is misidentified as 'MIRVISS'. We argue that the sampling module is considered to play the role of parallel attention. When the model is deployed with attention-based decoder, the worse performance may result from more difficult joint optimization of both sampling module and attention-based decoder.

B COMPARISON WITH TRADITIONAL TWO-STAGE TEXT SPOTTER

To validate the effectiveness of proposed single shot method SRSTS, we compare SRSTS with two-stage spotter by evaluating the number of trainable parameters, 'None' metric on Total-Text and recognition error rate when the IoU of the detected polygon and gt is greater than 0.5. ABCNet v2 which shows superiority over other two-stage spotters in terms of speed and accuracy is selected to be the counterpart. The detailed comparison is listed in Table 6. While SRSTS achieves better end-to-end performance, it also has the smaller model size, which reflects the advantages of SRSTS in dealing with spotting task. Benefiting from the absence of RoI operations, our SRSTS avoids the disadvantage of RoI cropping and the error propagation from text detection. To prove this point, we calculate the recognition error rate when the IoU of the detected polygon and groundtruth polygon is greater than 0.5. The error rates of SRSTS and ABCNet v2 are 17.6% and 21.2% respectively, which demonstrates the effectiveness of SRSTS.

Table 6: Comparison between SRSTS and ABCNet v2. The metrics used for comparison are the number of trainable parameters, 'None' metric on Total-Text and the recognition error rate on Total-Text when the IoU of the detected polygon and gt is greater than 0.5.

| Method    | Parameters | None  | Error rate |
|-----------|------------|-------|------------|
| ABCNet v2 | 47.7M      | 70.40 | 21.2%      |
| SRSTS     | 37.8M      | 78.80 | 17.6%      |