PAN++: Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text

Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Zhibo Yang, Tong Lu, Chunhua Shen

Abstract—Scene text detection and recognition have been well explored in the past few years. Despite the progress, efficient and accurate end-to-end spotting of arbitrarily-shaped text remains challenging. In this work, we propose an end-to-end text spotting framework, termed PAN++, which can efficiently detect and recognize text of arbitrary shapes in natural scenes. PAN++ is based on the kernel representation that reformulates a text line as a text kernel (central region) surrounded by peripheral pixels. By systematically comparing with existing scene text representations, we show that our kernel representation can not only describe arbitrarily-shaped text but also well distinguish adjacent text. Moreover, as a pixel-based representation, the kernel representation can be predicted by a single fully convolutional network, which is very friendly to real-time applications. Taking the advantages of the kernel representation, we design a series of components as follows: 1) a computationally efficient feature enhancement network composed of stacked Feature Pyramid Enhancement Modules (FPEMs); 2) a lightweight detection head cooperating with Pixel Aggregation (PA); and 3) an efficient attention-based recognition head with Masked RoI. Benefiting from the kernel representation and the tailored components, our method achieves high inference speed while maintaining competitive accuracy. Extensive experiments show the superiority of our method. For example, the proposed PAN++ achieves an end-to-end text spotting F-measure of 64.9 at 29.2 FPS on the Total-Text dataset, which significantly outperforms the previous best method. Code will be available at: git.io/PAN

Index Terms—End-to-End Text Spotting, Text Detection, Kernel Representation, Segmentation.

1 INTRODUCTION

Reading text in natural scenes is a fundamental task in numerous computer vision applications, such as text retrieval, office automation, and visual question answering. In virtue of the powerful representation of deep neural networks [1, 2], scene text detection and recognition has witnessed great progress in the past few years [3, 4, 5, 6, 7, 8]. Nonetheless, three main limitations still exist in these methods and hamper their deployment to real-world applications.

First, many works tackle text detection and recognition as separate tasks by focusing on either text detection or text recognition alone. For most existing text detectors [3, 4, 12], a convolutional neural network is first used to generate feature maps of an input image, and then rectangular or polygonal bounding boxes of scene texts are generated using a decoder. On the other side, text recognition methods [5, 6, 13] often conduct a sequential prediction network on top of image patches of text lines. To date, few methods explore the complementarity between these two tasks. As a result, computation overhead is introduced when assembling these standalone methods into a single scene text reading system.

Second, most end-to-end text spotters [7, 14, 15] are typically designed to read horizontal or oriented text lines. These methods assume that the layout of scene texts is straight, and

- W. Wang and T. Lu are with National Key Lab for Novel Software Technology, Nanjing University. Email: wangwenhai362@sml.ai, lutong@nju.edu.cn
- E. Xie is with The University of Hong Kong. Email: xieenze@hku.hk
- X. Li is with Nanjing University of Science and Technology. Email: xiang.li.implus@njust.edu.cn
- X. Liu and D. Liang are with SenseTime Group Limited. Email: {liuxuebo, liangding}@sensetime.com
- Z. Yang is with Alibaba Group. Email: zhibo.yzb@alibaba-inc.com
- C. Shen is with Monash University, Australia. Email: chunhua@me.com
- W. Wang and E. Xie contributed equally. T. Lu is the corresponding author.

Thus, both their detection and recognition components are developed upon this assumption. However, besides straight text lines, text lines with irregular character arrangements are very common in

Fig. 1. Comparisons of different scene text representations. (a) The quadrilateral representation [3, 7, 9] fail to locate curved text lines. (b) The pixel-wise representation [10] is not able to separate adjacent text lines. (c) Although the bounding-box-pixel (bbox-pixel) representation [7, 9, 11] can accommodate curved and adjacent text lines, it needs two-stage prediction, which can be computationally expensive. (d) Our kernel representation is one-stage and is robust to curved and adjacent text lines.
natural scenes, for example, texts in signboards and posters on streets. The methods dedicated to straight text lines would fail to correctly detect and recognize text lines of curved shapes as shown in Fig. 1(a).

Last, the efficiency of existing end-to-end text spotters [7], [8], [14], [15], [16] remains insufficient for real-world applications. Although a few recent methods [8], [16] have improved the accuracy of end-to-end arbitrarily-shaped text spotting, they suffer from low inference speed due to their heavy models or complicated pipelines. Thus, how to design an efficient and accurate end-to-end spotter for arbitrarily-shaped texts remains largely unsolved.

To address the above issues, built upon our preliminary results published in [4], [17], here we employ the kernel representation to describe a text line by a text kernel (central region) surrounded by peripheral pixels. This is inspired by systematically studying (dis)advantages of the existing scene text representations listed in Fig. 1. The quadrilateral representation [3], [7], [9] is the most commonly used one. As shown in Fig. 1(a), it is specifically designed for straight text lines and fails to provide a tight boundary for a curved text line. In contrast, the pixel-wise representation proposed in [10] is flexible enough to locate curved text lines. Note that with this representation, adjacent text lines can be conglutinated as shown in Fig. 1(b). Although the bounding-box-pixel (bbox-pixel) representation (see Fig. 1(c)) proposed in Mask R-CNN [18] can solve the problems mentioned above, it still follows “detect-then-segment” paradigm [18], which is often time-consuming. In contrast, as shown in Fig. 1(d), our kernel representation can effectively distinguish curved text lines lying closely, which is not inferior to the bbox-pixel representation. Besides, it can be predicted through a single convolutional network, which is conceptually simple and suitable for real-time applications.

By leveraging the advantages of the kernel representation, we further present an end-to-end arbitrarily-shaped text spotter, namely, PAN++ [4] which can achieve a good balance between accuracy and inference speed. PAN++ follows the pipeline shown in Fig. 1(h), which contains two main steps: 1) detecting text lines via a segmentation network; and 2) recognizing text content via an attention-based decoder.

To achieve high efficiency, we reduce the time cost of each step by the following four designs: 1) First, a lightweight backbone network (e.g., ResNet18 [19]) is employed for the segmentation network. However, features extracted by a lightweight backbone typically have small receptive fields and weak representation capabilities. 2) To remedy this defect, we propose a low-computation feature enhancement network, which is composed of stacked Feature Pyramid Enhancement Modules (FPEMs). Considering that FPEM is a U-shaped module built by separable convolutions (see Fig. 5), it can enhance multi-scales features with the minimal computation overhead. Moreover, FPEM is stackable, which allows us to compensate for the depth of the network by stacking FPEMs to the lightweight backbone. 3) To detect text lines, we design a simple detection head along with Pixel Aggregation (PA). The detection head predicts text regions, text kernels, and instance vectors, and PA assembles predictions of the network into complete text lines. 4) Finally, to recognize text content, we present a feature extractor termed Masked RoI and a lightweight recognition head that contains only two LSTM [20] layers and two multi-head attention layers [21]. Benefiting from the above designs, PAN++ achieves a high inference speed while keeping competitive accuracy.

Compared with our conference versions PSENet [4] and PAN [17], the major extension of PAN++ lies in the text recognition module and the end-to-end text spotting framework. In conference versions, we proposed the kernel representation and applied it to text detection. Different from the predecessors, we extend the text detector to an end-to-end text spotter that can fast detect and recognize arbitrarily-shaped text lines. To this end, we reconstruct the overall architecture of PAN++, and carefully integrate a tailored feature extractor (i.e., Masked RoI) and a lightweight text recognition head in the architecture.

Besides, we also improve the text detection module when inheriting it from the conference version [17]. 1) We systematically compare our kernel representation with other existing text representations. 2) We simplify the FPEM by combining the original FPEM proposed in [17] and FFM [17] into a single module, which is more effective. 3) We make PA be aware of background elements by adding a background item to the discrimination loss. Combining these improvements, we upgrade the original text detector into an efficient end-to-end arbitrarily-shaped text spotter (i.e., PAN++). To show the effectiveness of our method, we conduct extensive experiments on four challenging benchmark datasets, namely, Total-Text [23], CTW1500 [24], ICDAR 2015 [25], MSRA-TD500 [26], and RCTW-17 [61]. Note that, Total-Text and CTW1500 are datasets created for curved text detection. As shown in Fig. 2 on the Total-Text dataset, the end-to-end text spotting F-measure of “PAN++ 736” reaches 68.6%, which is 4.4 points higher than ABCNet [22], while its inference speed is faster. “PAN++ 512” executes faster than counterparts by over 11 FPS while keeping a competitive end-to-end text spotting F-measure.

In summary, our main contributions are four-fold.

- We provide the definition of the kernel representation, and
systematically compare it with other widely-used text representations, showing that our kernel representation is conceptually simple and flexible, and friendly to real-time applications.

- Based on the kernel representation, we propose a new framework for end-to-end arbitrarily-shaped text spotting, termed PAN++, which achieves an excellent balance between accuracy and inference speed.

- We design and improve a series of efficient components tailored to our framework, including a feature enhancement network consisting of stacked Feature Pyramid Enhancement Modules (FPEMs), a detection head with Pixel Aggregation (PA), a feature extractor (Masked RoI), and an lightweight attention-based recognition head.

- The proposed PAN++ achieves the state-of-the-art performance on curved text benchmarks, simultaneously keeping a fast inference speed. It is notable that PAN++ (S: 512) yields an end-to-end F-measure of 64.9% at 29.2 FPS on the TotalText dataset, which is 11 FPS faster than the previous best method with competitive accuracy.

## 2 RELATED WORK

We briefly review some work most relevant to ours.

### 2.1 End-to-End Text Spotting

End-to-end text spotters aim to detect and recognize text lines simultaneously in a unified network. In the past years, the emergence of deep learning has greatly improved the performance of text spotting. As shown in Fig. 3, we summarize representative text spotters relevant to ours in the deep learning era. These methods can be roughly divided into two categories: 1) regular text spotter, and 2) arbitrarily-shaped text spotter.

#### 2.1.1 Regular Text Spotter

Inspired by Faster R-CNN [27], Li et al. [14] proposed the first framework for horizontal text spotting, which contains a text proposal network for text detection and a CTC-based method for text recognition, as shown in Fig. 3(a). Meanwhile, Busta et al. [28] designed a similar framework to the work in [14]. But its detection head works better for both horizontal and multi-oriented text instances. NguyenVan et al. [29] developed an end-to-end text reading framework by incorporating a pooling-based scene text proposal technique, where false alarms elimination and words recognition are performed simultaneously. Subsequently, Liu et al. [7] proposed a multi-oriented text spotter shown in Fig. 3(b), termed FOTS, which is equipped with a new RoI extractor (i.e., RoIRotate) to extract the features of quadrilateral text instances. A similar framework is also developed by He et al. [15], whose recognition head is implemented by an attention-based sequence-to-sequence decoder. Although these methods have achieved good performance across straight text benchmarks (e.g., IC15 [25] and MSRA-TD500 [26]), they fail to detect and recognize text lines of curved shapes (see Fig. 3(a)).

#### 2.1.2 Arbitrarily-Shaped Text Spotter

As illustrated in Fig. 3(c), Liao et al. [8] proposed Mask TextSpotter (MTS) by adding character-level supervision to Mask R-CNN [18] to detect and recognize text lines and characters simultaneously. This is probably the first arbitrarily-shaped text spotter. Note that character-level annotations are not always available. The improved version [30] of Mask TextSpotter removed the dependence of character-level annotations. CharNet in [31] is the first one-stage arbitrarily-shaped text spotter as shown in Fig. 3(d). Similar to Mask TextSpotter, it requires character-level annotations for training. Feng et al. [32] proposed TextDragon as shown in Fig. 3(e), which contains a novel RoI extractor (i.e., RoISlide) to represent a text line by features of multiple text segments. Qin et al. [16] designed an arbitrarily-shaped text spotter as shown in Fig. 3(f), where an attention-based text recognition head was added on Mask R-CNN to simultaneously detect and recognize irregular text lines. Recently, Liu et al. [22] formulated irregular text instances with parameterized Bezier curves, and proposed
a regression-based arbitrarily-shaped text spotter ABCNet (see Fig. 3(g)). Note that pixel-based (mask) representation used in the previous methods \[4, 8, 16\] may be more flexible and reliable than the Bezier curve representation due to the fact that the latter one can only represent a series of curves under specific constraints.

2.1.3 Real-time Arbitrarily-Shaped Text Spotter

In \[7\], Liu et al. proposed a fast text spotter FOTS to detect and recognize multi-oriented texts. However, FOTS cannot process curved text lines. Most methods \[8, 16, 30, 31, 32\] mentioned in Sec. 2.1.2 have achieved high accuracies on arbitrarily-shaped text spotting, but the inference speed is rarely addressed in those methods.

This inspires us to design an improved pixel-based representation and develop an effective end-to-end text spotting framework based on it in this work.

2.2 Scene Text Detection

Scene text detection is a critical part of text spotting systems. Recently, methods based on deep learning have become the mainstream of text detection. Tian et al. \[12\] and Liao et al. \[33\] successfully adapted object detection frameworks for text detection, and achieved good performance on horizontal text datasets. Then researchers took orientations of text instances into consideration and developed different methods \[3, 9, 34\] to detect multi-oriented text instances. Most of these methods make text detection by regressing quadrilateral bounding boxes, failing to detect curved text lines. To detect curved text lines, some early works \[34, 53\] followed a bottom-up framework to first detect characters or text fragments and then link them. Recently, some researchers designed segmentation-based methods \[4, 17, 36\] or combined the advantages of segmentation and box/point regression \[13, 37, 38\], achieving excellent performances on curved text detection benchmarks. These methods were designed to solve text detection alone and typically text recognition is achieved using a separate model.

In this work, we not only improve the performance of text detection by summarizing the advantages of our conference versions (i.e., PSENet \[4\] and PAN \[17\]), but also carefully integrate a recognition module in our framework to build a unified model for text spotting.

2.3 Scene Text Recognition

Scene text recognition is the last step in text spotting systems, which aims to decode sequence of characters from the visual input. Existing text recognizers can be roughly split into three categories, namely, 1) character-based text recognizers, 2) CTC-based text recognizers, and 3) attention-based text recognizers. Character-based text recognizers \[41\] mostly first detect individual characters and then group them into words. CTC-based methods \[5, 42, 43, 44\] often stack RNN on top of CNN to capture long-range sequence features. These methods are trained with CTC loss \[39\], and their the final prediction is produced by removing duplicated outputs in the testing phase. In attention-based methods \[6, 13, 45, 46\], RNN or Transformer \[21\] is used to predict a character at each step based on the features from attention mechanisms and previous steps. The process will stop when it predicts the end-of-sequence (EOS) or reaches the maximum number of steps. To recognition irregular text lines, Shi et al. \[6\] and Zhan et al. \[47\] used a spatial rectification network to rectify irregularly-shaped text lines into straight ones before recognition. Besides that, some recent works \[13, 45\] applied the attention mechanism on two-dimensional feature maps to recognize texts with irregular shapes.

Inspired by \[13\] and \[21\], we design a lightweight attention-based recognition head for irregular text recognition by utilizing multi-head attention to fuse visual and temporal features from CNN and RNN, respectively.

3 Proposed Method

3.1 Kernel Representation

As presented in Fig. 1(d), the kernel representation formulates a text line as a text kernel surrounded by peripheral pixels. In other words, for each given text line, we first locate it through the text kernel (the central region of the text line). Then, we recover the complete shape of the text line by involving text pixels around the text kernel.

In general, there are four advantages of our kernel representation. 1) As a pixel-based representation, it is flexible enough to represent text lines of arbitrary shapes; 2) Because there are large geometrical margins among text kernels, the kernel representation can accurately distinguish adjacent text instances, solving the conglutination problem caused by adjacent text lines shown in Fig. 1(b); 3) The kernel representation is totally pixel-based, meaning that it can be easily predicted by a single fully convolutional network, which is friendly to real-time applications. 4) The label of the kernel representation can be simply generated without additional annotations, as shown in Fig. 10.

3.2 Overall Architecture

Based on the kernel representation, we develop an efficient text spotting framework termed PAN++, whose architecture is illustrated in Fig. 4. For high efficiency, the backbone network (CNN) should be lightweight, such as ResNet18. However, the features produced by the lightweight backbone network often have small receptive fields and weak representation capabilities. To handle this problem, we propose a feature enhancement network that can refine the features efficiently. It consists of several stacked Feature Pyramid Enhancement Modules (FPEMs). As shown in Fig. 5, FPEM is stackable and computationally efficient, which can be attached behind the backbone network and makes features deeper and more expressive than before. For text detection, we propose a simple detection head with Pixel Aggregation (PA), which predicts: 1) text regions to describe the complete shapes of text lines, 2) text kernels to distinguish different text lines, and 3) instance vectors to recover complete text lines from text kernels, and then these predictions are combined into the text line by PA. For text recognition, Masked RoI is employed to extracted feature patches of text lines, and an attention-based recognition head is proposed to recognize text content.

In the inference phase, we firstly feed an input image of a size of \(H \times W \times 3\) to a lightweight backbone network (e.g., ResNet18 \[19\]). Four feature maps (see Fig. 4(b)) are generated by conv2, conv3, conv4, and conv5 stages of the backbone network, whose resolutions are 1/4, 1/8, 1/16 and 1/32 compared with the input image, respectively. Then, we reduce the channel number of each feature map to 128 via \(1 \times 1\) convolutions, and obtain a thin feature pyramid \(F_r\) (see Fig. 4(c)). The feature pyramid is enhanced by \(N_{stk}\) stacked FPEMs. After obtaining enhanced
feature pyramid $F_e$ (see Fig. 4 (d)), we upsample and concatenate the feature maps of feature pyramid $F_f$ into the final feature map for follow-up text detection and recognition. The final feature map is termed $F_f$ (see Fig. 4 (e)), whose size is $H/4 \times W/4 \times 512$. Next, based on the feature map $F_f$, the detection head predicts text regions, text kernels, and instance vectors, and then Pixel Aggregation (PA) assembles them into complete text lines (see Fig. 4 (i)). During text recognition, we first reduce the channel number of feature map $F_f$ to 128 by a $3 \times 3$ convolution and apply Masked RoI to extract the feature patches according to the predicted text lines. Finally, using the extracted feature patches as input, the recognition head recognizes the text content in each patch, as shown in Fig. 4 (j).

During training, we use loss functions $L_{tex}, L_{ker}, L_{agg},$ and $L_{dis}$ to optimize the predicted text regions, text kernels and instance vectors, respectively. At the same time, the loss function $L_{rec}$ is applied to optimize the prediction of the recognition head. To maintain the consistency of recognition features, we use ground-truth bounding boxes to extract feature patches in the training phase.

### 3.3 Feature Pyramid Enhancement Module

The Feature Pyramid Enhancement Module (FPEM) is the basic unit of the feature enhancement network, which is U-shaped as presented in Fig. 5. FPEM consists of two phases, including up-scale enhancement and down-scale enhancement phases. The up-scale enhancement is applied to the input feature pyramid, which enhances the input feature maps with strides of 32, 16, 8, and 4 pixels iteratively. In the down-scale enhancement phase, the input is the feature pyramid generated by up-scale enhancement, and the enhancement is conducted from 4-stride to 32-stride iteratively. Finally, the output feature pyramid is the element-wise addition result of the input feature pyramid and the feature pyramid produced by the down-scale enhancement.

As shown in the dashed boxes in Fig. 5, we employ the separable convolution [48] (a $3 \times 3$ depthwise convolution followed by a $1 \times 1$ projection) instead of the regular convolution to implement the join part $\oplus$ of FPEM. Therefore, FPEM is capable of enlarging the receptive field (thanks to $3 \times 3$ depthwise convolutions) and deepening the network (thanks to $1 \times 1$ projections) with a small computation overhead.

Similar to FPN [51], FPEM can enhance multi-scale feature maps by fusing the low-level and high-level information. Beyond that, there are two other advantages of FPEM. 1) FPEM is a stackable module. With the increment of stack number $N_{stk}$, the feature maps will be more adequately integrated, and their receptive fields will become larger. 2) FPEM is computationally efficient. Benefiting from separable convolutions, FPEM only involve marginal computational overhead. As reported in Table 8, the model equipped with FPEMs runs 8.7 FPS faster than the model with FPN on the IC15 dataset, while keeping a higher end-to-end text spotting F-measure.

### 3.4 Text Detection

#### 3.4.1 Detection Head

The detection head of PAN++ consists of only two convolutions, as shown in Fig. 6. The outputs of the detection head are text regions, text kernels, and instance vectors. 1) Text regions keep the complete shape of text lines, but closely lying text lines are often
Text Kernel

![Text Kernel](https://example.com/text_kernel.png)

Fig. 6. Details of the detection head. “Conv” and “BN” represent regular convolution [49] and Batch Normalization [50], respectively.

**3.4.2 Pixel Aggregation**

Pixel aggregation (PA) is applied to optimize instance vectors and assemble the outputs of the detection head. We borrow the idea of clustering to design PA. As shown in Fig. 7, if we treat different text lines as different clusters, their text kernels are cluster centers, and the pixels inside text regions are the samples to be clustered.

Naturally, to cluster pixels to the corresponding text kernels (see the green arrows in Fig. 7), the distance between the pixel and the text kernel inside the same text lines should be minimized. In the training phase, we use an aggregation loss $\mathcal{L}_{agg}$ formulated as Eq. (1) to implement this rule.

$$
\mathcal{L}_{agg} = \frac{1}{N} \sum_{i=1}^{N} \sum_{p \in T_i} \mathcal{D}_1(p, K_i),
$$

(1)

$$
\mathcal{D}_1(p, K_i) = \ln(\mathcal{R}(\|\mathcal{F}(p) - \mathcal{G}(K_i)\| - \delta_{agg})^2 + 1).
$$

(2)

Generally, the aggregation loss $\mathcal{L}_{agg}$ is employed to pull the text pixel to the target text kernel. The symbols in Eqs. (1) and (2) are defined as follows:

- $N$: the number of text lines.
- $T_i$: the text region of the $i$th text line.
- $K_i$: the text kernel of the text line $T_i$.
- $\mathcal{D}_1(p, K_i)$: the distance between the text pixel $p$ and the text kernel $K_i$.
- $\mathcal{R}(\cdot)$: the ReLU function, ensuring the output is non-negative.

![Illustration of Pixel Aggregation (PA)](https://example.com/pixel_aggregation.png)

Fig. 7. Illustration of Pixel Aggregation (PA). Green arrows represent aggregation loss $\mathcal{L}_{agg}$. Red arrows represent discrimination loss $\mathcal{L}_{dis}$.

overlapping (see Fig. 4(f)). 2) In contrast, text kernels can easily distinguish adjacent text instances, but they lose the complete shape of text lines (see Fig. 4(g)). 3) The instance vector is a high-dimensional representation that contains instance information of each pixel. In other words, pixels belonging to the same text line tend to have similar instance vectors (see Fig. 4(h)). Therefore, instance vectors can direct the pixels in text regions towards the corresponding text kernel, which successfully combines the advantages of text regions and text kernels.

The three outputs complement each other and make it possible to describe text lines of arbitrary shapes.

The goal of the discrimination loss $\mathcal{L}_{dis}$ is to push the text kernels and the background away from each other. The following is the definition of the symbols in Eqs. (3), (4), and (5).

- $\mathcal{F}(p)$: the instance vector of the pixel $p$.
- $\mathcal{G}(K_i)$: the instance vector of the text kernel $K_i$, which can be calculated by: $\mathcal{G}(K_i) = \sum_{p \in K_i} \mathcal{F}(p) / |K_i|$, where $|K_i|$ is the number of pixels in $K_i$.
- $\delta_{agg}$: a constant, which is set to 0.5 experimentally.

Besides that, the cluster centers need to keep discrimination correspondingly. As a result, the instance vector of the text kernel should keep a certain distance from other text kernels and the background, as shown in the red arrows in Fig. 7. This rule can be summarized as a discrimination loss $\mathcal{L}_{dis}$ as follows:

$$
\mathcal{L}_{dis} = \frac{1}{N^2} \sum_{i=1}^{N} \mathcal{D}_b(K_i) + \sum_{j=1}^{N} \mathcal{D}_2(K_i, K_j)),
$$

(3)

$$
\mathcal{D}_b(K_i) = \frac{1}{|B|} \sum_{p \in B} \ln(\mathcal{R}(\delta_{dis} - \|\mathcal{F}(p) - \mathcal{G}(K_i)\|)^2 + 1),
$$

(4)

$$
\mathcal{D}_2(K_i, K_j) = \ln(\mathcal{R}(\delta_{dis} - \|\mathcal{G}(K_i) - \mathcal{G}(K_j)\|)^2 + 1).
$$

(5)

3.5 Text Recognition

3.5.1 Masked RoI

Masked RoI is an RoI extractor used to extract the fixed-size feature patches for arbitrarily-shaped text lines. It contains four steps: 1) calculating the minimal upright bounding rectangle containing the target text line; 2) extracting the feature patch within the upright bounding rectangle. 3) filtering noise features by multiplying the feature patch by a binary mask, where the weight
outside the target text line is 0. 4) resizing the feature patch to a fixed size. These steps can be summarized as Eqn. (6).

\[ F_{roi} = Resize_{h \times w}(Crop(F_f, Rect(m_i)) \circ m_i), \]  

(6)

where \( m_i \) is the binary mask of the \( i \)th text line. \( Rect(\cdot) \) refers to the upright bounding rectangle of a binary mask. \( Crop(F_f, \cdot) \) is the operation of cropping the feature patch inside an upright RoI region from the feature map \( F_f \). \( \circ \) means element-wise multiplication. \( Resize_{h \times w}(\cdot) \) represents the operation of resizing the feature map’s size to \( h \times w \) via bilinear interpolation. In all experiments, \( h \) and \( w \) are set to 8 and 32 pixels, respectively.

The proposed Masked RoI has two benefits: 1) The binary mask of the target text line can eliminate the noise features caused by the background or other text lines, so as to accurately extract the features of arbitrarily-shaped text lines. 2) Masked RoI skips the spatial rectification step (e.g., STN in ASTER [6]), reducing the time cost of feature extraction.

Although our Masked RoI is somewhat analogous to RoI Masking proposed in [16], there is a major distinction between them. In RoI Masking [16], the mask used to filter noise features is an instance-wise attention map, whose weight is soft (i.e., a floating number). This approach is not suitable for our PAN++, which cannot predict instance-wise soft masks of text lines. Differently, our Masked RoI removes noise by the binary mask, which can be easily generated by PA. Moreover, the noise weight in the binary mask is 0. Therefore, the proposed Masked RoI can remove noise more thoroughly than RoI Masking.

3.5.2 Recognition Head

Our recognition head is a seq2seq model equipped with multi-head attention [21]. As shown in Fig. 9 it is composed of a starter and a decoder.

The starter is proposed to find the start of the string (SOS), which is not necessarily the left-most region of an arbitrarily-shaped text line. The starter only contains a linear transformation (embedding layer) \( \mathcal{E}_1 \) and a multi-head attention layer \( \mathcal{A}_1 \). As described in Eqn. (7), the embedding layer \( \mathcal{E}_1 \) transfer the SOS symbol (one-hot) to a 128-dim vector, and then this vector is fed to the multi-head attention layer \( \mathcal{A}_1 \) along with the flattened feature patch \( F_{roi} \), producing SOS’s feature vector \( \mathbf{f}_s \), which is also the input of the decoder at the initial time step.

\[ \mathbf{f}_s = \mathcal{A}_1(\mathcal{E}_1(\text{"SOS"}), F_{roi}) \]  

(7)

The decoder is made up of only two LSTM layers and one multi-head attention layer \( \mathcal{A}_2 \). At time step 0 (initial time step), the decoder takes SOS’s feature vector \( \mathbf{f}_s \) and zero LSTM initial states as input. At time step 1, we input the hidden state \( \mathbf{h}_0 \) of time step 0 into LSTM along with the SOS symbol, and predict the output symbol \( y_1 \) of time step 1. After that, the output symbol of the previous step is fed into LSTM until predicting the end of the string (EOS) token. We formulate this process as follows:

\[ y_t = \text{argmax}(FC(A_2(h_t, F_{roi}))), \quad 1 \leq t \leq T, \]  

(8)

\[ h_0 = \text{LSTM}(\mathbf{f}_s, \mathbf{0}), \]  

(9)

\[ h_t = \text{LSTM}(\mathcal{E}_2(y_{t-1}), h_{t-1}). \]  

(10)

Here, \( \mathbf{0} \) represents zero states. All input symbols (including SOS and EOS) are represented by one-hot vectors, followed by an embedding layer \( \mathcal{E}_2 \). \( y_t \) is the output symbol of the time step \( t \). During training, the input symbols of the decoder are characters in the ground-truth sequence.

Our recognition head is lightweight while maintaining good accuracy. Our recognition head has no encoder, and its decoder only contains two LSTM layers and an attention layer. Moreover, our attention layer is based on multi-head attention, which can effectively fuse temporal (LSTM) features and visual (CNN) features. Besides that, the computational cost of multi-head attention is lower than the common attention module used in [13].

3.6 Label Generation

As presented in Fig. 4(f)(g), our method predicts the masks of text regions and text kernels. Therefore, the corresponding mask label is required for training.

For text regions, their mask labels can be directly generated by filling ground-truth bounding boxes, where the text pixel is 1 and the non-text pixel is 0. Here, we denote the mask label of the text region as \( G_{rer} \).

For text kernels, we generate their mask labels by shrinking the original bounding boxes by a certain margin and filling them. As shown in Fig. 10 the polygon with a blue border denotes the original bounding box of a text line. To generate the label of the text kernel, we firstly utilize the Vatti clipping algorithm [52] to shrink the original bounding box \( b_d \) by \( m \) pixels and get shrunk bounding box \( b_k \) (see Fig. 10(a)). After that, the shrunk bounding box \( b_k \) is transferred into a binary mask, which is the label of the text kernel. For convenience, we denote the mask labels of text kernels as \( G_{ker} \).

During the label generation of text kernels, if we consider a shrinking rate \( r \in [0, 1) \), the shrinking margin \( m \) between the original bounding box \( b_d \) and the shrunk bounding box \( b_k \) can be calculated as:

\[ m = \frac{\text{Area}(b_d) \times (1 - r^2)}{\text{Perimeter}(b_d)}. \]  

(11)

Here, \( \text{Area}(\cdot) \) is the function of computing the area of a polygon. \( \text{Perimeter}(\cdot) \) is the function of computing the perimeter of a polygon. In the training phase, we can control the scale of target text kernels by adjusting the shrinking rate \( r \).

3.7 Loss Function

Our loss function can be formulated as Eqn. (12).

\[ L = L_{\text{det}} + L_{\text{rec}}, \]  

(12)

where \( L_{\text{det}} \) and \( L_{\text{rec}} \) are the loss functions for text detection and recognition, respectively.
we only take text pixels into consideration when calculating $L$. To ignore simple non-text pixels when calculating $L$, we also adopt Online Hard Example Mining (OHEM) \[54\] to optimize the segmentation result of text regions $P_{\text{tex}}$ and of text kernels $P_{\text{ker}}$. Therefore, $L_{\text{tex}}$ and $L_{\text{ker}}$ are formulated as:

$$L_{\text{tex}} = 1 - \frac{2 \sum_{i} P_{\text{tex}}(i) G_{\text{tex}}(i)}{\sum_{i} P_{\text{tex}}(i)^2 + \sum_{i} G_{\text{tex}}(i)^2},$$

$$L_{\text{ker}} = 1 - \frac{2 \sum_{i} P_{\text{ker}}(i) G_{\text{ker}}(i)}{\sum_{i} P_{\text{ker}}(i)^2 + \sum_{i} G_{\text{ker}}(i)^2}.$$ 

Here, $P_{\text{tex}}(i)$ and $G_{\text{tex}}(i)$ refer to the value of the $i$th pixel in the segmentation result and the ground truth of the text regions, respectively. Similarly, $P_{\text{ker}}(i)$ and $G_{\text{ker}}(i)$ means the $i$th pixel value in the prediction and the ground truth of text kernels, respectively. We also adopt Online Hard Example Mining (OHEM) \[54\] to ignore simple non-text pixels when calculating $L_{\text{tex}}$. Note that, we only take text pixels into consideration when calculating $L_{\text{ker}}$, $L_{\text{agg}}$, and $L_{\text{dis}}$.

The loss function $L_{\text{rec}}$ of text recognition is:

$$L_{\text{rec}} = \frac{1}{|w|} \sum_{i=0}^{|w|} \text{CrossEntropy}(y_i, w_i),$$

where $w$ is the ground-truth transcription (text content) containing EOS symbol. $|w|$ denotes the number of characters in the transcription. $w_i$ is the $i$th character in the transcription.

4 EXPERIMENT

To validate the effectiveness of the proposed PAN++, we evaluate it on two challenging tasks: 1) text detection and 2) end-to-end text spotting.

4.1 Datasets

4.1.1 Curved Text Datasets

**Total-Text** \[23\] is a dataset for arbitrarily-shaped text detection and spotting, containing horizontal, multi-oriented, and curved text lines. This dataset consists of 1,255 training images and 300 testing images, all of which are annotated with polygons and transcriptions at the word level.

**CTW1500** \[24\] is a widely used dataset for arbitrarily-shaped text detection. It has 1,000 training images and 500 testing images. Different from the Total-Text dataset, the images in this dataset are labeled at the text line level. The text lines in the CTW1500 dataset are long and annotated by 14-points polygons.

4.1.2 Straight Text Datasets

**ICDAR 2015** (IC15) \[25\] is a commonly used dataset for end-to-end text detection and recognition. It contains a total of 1,500 images, 1,000 of which are used for training and the remaining are for testing. In this dataset, text lines are annotated by quadrangles and transcriptions at the word level.

**MSRA-TD500** \[58\] includes 300 training images and 200 test images with text line level annotations. It contains multi-lingual, multi-oriented, and long text lines, and all of them are labeled at the text line level. Because its training set is small, we follow the previous works \[3, 37, 59\] to include the 400 images of HUST-TR400 \[60\] as training data.

**RCTW-17** \[61\] is a competition on reading Chinese Text in images. It provides a large-scale dataset that consists of 12,000 images, where 8,346 for training and the rest for testing. Every image in this dataset is annotated with text line quadrilaterals and text transcripts.

4.1.3 Datasets for Pre-training or Jointly Training

**SynthText** \[62\] is a large-scale synthetically generated dataset containing 800K synthetic images. There are a large number of multi-oriented text lines, which are annotated at word-level and character-level rotated bounding boxes, as well as corresponding transcriptions. Following \[34, 37, 59\], this dataset is used for pre-training or jointly training models of PAN++.

**COCO-Text** \[63\] contains 63,686 images and three versions of the annotations (V1.1, V1.4, and V2.0). In V2.0, there are 239,506 annotated text lines, which are labeled with word-level polygons and transcriptions. This dataset is usually used for pre-training or jointly training in previous text spotting methods \[16, 22\].

**ICDAR 2017 MLT** (IC17-MLT) \[64\] is a multi-language scene text dataset. It consists of 7,200 training images, 1,800 validation images, and 9,000 test images. The dataset is composed of natural scene images with nine languages, and these images are labeled with word-level quadrangles and transcriptions. Following \[16, 22\], we use the English samples in the training set to pre-train or jointly train our text spotting models.

4.2 Experiments on Curved Text Datasets

4.2.1 Experiment Settings

To test the performance of PAN++ on arbitrarily-shaped text, we conduct experiments on Total-Text and CTW1500 and compare it with previous start-of-the-art methods. In these experiments, we use the ResNet \[2\] pre-trained on ImageNet \[65\] as the backbone network of our method. In addition, we set the dimension of the instance vector to 4, the negative-positive ratio of OHEM to 3, the shrinking rate $r$ of the text kernel to 0.7, and the distance threshold $d$ of PA to 3.

Following the common practices \[3, 16, 28, 22\], we ignore the blurred text regions labeled as "DO NOT CARE" during training, and apply random scale, random horizontal flip, random rotation, and random crop on training images. All models are optimized by using ADAM \[66\] optimizer with a batch size of 16 on 4 GPUs. The initial learning rate is set to $1 \times 10^{-3}$. Similar
to [67], we use the “poly” learning rate strategy in which the initial rate is multiplied by \((1 - \frac{\text{iter}}{\text{max iter}})^\text{power}\), and the power is set to 0.9 in all experiments. In the testing phase, we resize the input image to different scales and report the performance on text detection and end-to-end text spotting tasks. All results are tested with a batch size of 1 on a V100 GPU and a 2.20GHz CPU in a single thread unless otherwise stated.

In the text detection task, to make fair comparisons, we train PAN++ (without the recognition head) with two widely used strategies as follows: 1) following [3], [17], [68] to train models without external text datasets; 2) following [4], [17], [34] to fine-tune models pre-trained on external text dataset (e.g., SynthText and IC17-MLT). In the first strategy, we train our method for 36K iterations. In the second one, we first pre-train models on the external text dataset for 50K iterations and then fine-tune the pre-trained models for 36K iterations.

In the text spotting task, the training strategies used in previous works [7], [8], [16], [22] can also be divided into two categories: 1) fine-tuning models pre-trained on external text datasets, and 2) jointly training models on multiple text datasets. For fair comparisons, we train our end-to-end text spotting models using the two strategies respectively. Specifically, when using the fine-tuning strategy, we first pre-train models on the training images in the SynthText, COCO-Text, and IC17-MLT datasets for 150K iterations, and then we fine-tune the pre-trained models on Total-Text or CTW1500 for 7K iterations. For the jointly training strategy, we train our models for 150K iterations on the joint dataset that includes all training samples in SynthText, COCO-Text, IC17-MLT, Total-Text and IC15.

### 4.2.2 Curved Text Detection Results

As shown in Table 1 without pre-training on external text datasets, PAN++ with the shorter side being 320 pixels yields an F-measure of 79.5% on CTW1500, surpassing most of the counterparts, including some methods (e.g., TextSnake [37]) pre-trained on SynthText (79.5% vs. 78.4%). Moreover, our inference speed is 8 times faster than TextSnake and 4 times faster than EAST [3] (the fastest method before). It notable that when fine-tuning the model pre-trained on SynthText, PAN++ (S: 640) achieves the best F-measure of 84.0%, outperforming all counterparts while still keeping the fastest inference speed (36 FPS).

Similar results are also on the Total-Text dataset. When the shorter side is 320 pixels, the inference speed of our method reaches 84.9 FPS, which is at least 6 times faster than previous methods, while the F-measure is still very competitive (80.0%). The best F-measure of our method is 85.3%, surpassing current state-of-the-art methods over 1.7 points. Meanwhile, its speed is close to 40 FPS, which is still 3 times faster than the second-fastest TextSnake [37].

These results prove the superiority of the proposed PAN++ (detection only) in arbitrarily-shaped text detection, in terms of accuracy and speed. We present some qualitative curved text detection results in Fig. 11(a)(b), which demonstrates that the detection components of PAN++ can elegantly locate text lines with complex shapes.

### 4.2.3 Curved Text Spotting Results

End-to-end text spotting results on the Total-Text dataset are reported in Table 2. Compared with the previous fastest ABC-Net [22], our method with shorter side being 640 pixels runs 1.3 times faster, while our end-to-end text spotting F-measure is 2.2 points higher (66.4% vs. 64.2%). When adopting a larger input scale or jointly training strategy, the performance of our method can be further improved. With low-resolution input (the shorter side being 512 pixels), the inference speed of our method reaches 29.2, at least 11 FPS faster than previous methods. At the same time, its end-to-end text spotting F-measure is 64.9%, which is higher than most counterparts. It is notable that PAN++ achieves the best end-to-end text spotting F-measure of 68.6%, which is much better than the second-best method (Qin et al. [16]) and our speed is 4 times faster (21.1 FPS vs. 4.8 FPS).

These results indicate that the proposed method achieves state-of-the-art performance in the text spotting task, significantly outperforming existing methods, especially in the inference speed. Some qualitative results of end-to-end curved text spotting are shown in Fig. 12(a).

### Table 1

Text detection results on Total-Text and CTW1500. “*” indicates the results from [37]. “†” indicates the results from [24]. “Scale” denotes the scale of the testing image, where “L.” means the longer side is fixed, and “S.” means the shorter side is fixed. Note that, we use a 1080Ti GPU to test the inference speed of PAN++ (detection only) for fair comparisons. Precision (%), Recall (%) and F-measure (%) are reported here.

| Method          | Scale | Backbone | External Dataset | Precision | Recall | F-measure | FPS | Precision | Recall | F-measure | FPS |
|-----------------|-------|----------|------------------|-----------|--------|----------|-----|-----------|--------|----------|-----|
| CTPN            | 600   | VGG16    | ×                 | -         | -      | -        | -   | 60.4†     | 53.8†  | 56.9†    | 7.1 |
| SegLink         | 768×1280 | VGG16   | ×                 | 30.3†     | 23.8†  | 26.7†    | -   | 42.3†     | 40.0†  | 40.8†    | 10.7|
| EAST            | 720×1280 | VGG16   | ×                 | 50.0†     | 36.2†  | 42.0†    | -   | 78.7†     | 49.1†  | 60.4†    | 21.2|
| DeconvNet       | 224×224 | VGG16   | ×                 | 33.0      | 40.0   | 36.0     | -   | -         | -      | -        | -   |
| CTD+TLOC        | 640   | ResNet50 | ×                 | -         | -      | -        | -   | 77.4      | 69.8   | 73.4     | 13.3|
| PAN++ (ours)    | 320   | ResNet18 | ×                 | 82.6      | 72.9   | 71.4     | 84.9| 82.9      | 76.3   | 79.5     | 84.2|
|                 | 512   | ResNet18 | ×                 | 87.1      | 79.0   | 82.8     | 53.1| 85.2      | 80.3   | 82.7     | 52.7|
|                 | 640   | ResNet18 | ×                 | 89.2      | 80.3   | 84.5     | 38.3| 85.2      | 81.1   | 83.1     | 36.0|
| TextSnake       | 512×512 | VGG16   | SynthText         | 82.7      | 74.5   | 78.4     | 12.4| 67.9      | 85.3   | 75.6     | -   |
| TextField       | 768×768 | VGG16   | SynthText         | 81.2      | 79.9   | 80.6     | -   | 83.0      | 79.8   | 81.4     | -   |
| CRAFT           | 1280  | VGG16    | SynthText         | 87.6      | 79.9   | 83.6     | 4.8 | 86.0      | 81.1   | 83.5     | 7.6 |
| SPCNet          | 800   | ResNet50 | IC17-MLT          | 88.6      | 75.7   | 81.6     | 4.4 | 89.2      | 69.6   | 78.4     | -   |
| PAN++ (ours)    | 320   | ResNet18 | SynthText         | 85.6      | 75.1   | 80.0     | 84.9| 84.5      | 76.5   | 80.3     | 84.2|
|                 | 512   | ResNet18 | SynthText         | 88.4      | 80.5   | 84.2     | 53.1| 86.9      | 80.4   | 83.5     | 52.7|
|                 | 640   | ResNet18 | SynthText         | 89.9†     | 81.0†  | 85.3†    | 38.3| 87.1†     | 81.1†  | 84.0†    | 36.0|


4.3 Experiments on Straight Text Datasets

4.3.1 Experiment Settings

We also evaluate the performance of PAN++ on straight text datasets, including IC15 and MSRA-TD500. IC15 is a representative benchmark for the text detection task and the text spotting task, and MSRA-TD500 is a dataset annotated at the text line level and often used to benchmark long text detection. During training, the shrinking rate $r$ on IC15 and MSRA-TD500 are set to 0.5 and 0.7, respectively. Other model settings, data augmentations, and training settings are the same as those in Sec. 4.2. In the testing phase, we rescale the shorter side of the input image to 736 or 896 pixels for fair comparisons with previous methods.

4.3.2 Straight Text Detection Results

Straight text detection results are presented in Table 3. On the IC15 dataset, PAN++ (the shorter side being 736 pixels) yields an F-measure of 81.2% at 28.2 FPS without pre-training on external text dataset. Compared with the previous fastest method EAST [3], our method surpasses EAST by 3.0 points in F-measure, while our method is twice faster. Because IC15 contains many small texts, previous methods [11], [57], [72] usually employ high-resolution input and a heavy backbone to ensure detection accuracy. With these setting, the best F-measure of our method reaches 87.5%, which is higher than most counterparts.

In conclusion, the proposed PAN++ (detection only) can also accurately detect straight text lines. Some qualitative results of straight text detection are shown in Fig. 11(c)(d).

4.3.3 Straight Text Spotting Results

Table 4 summarizes end-to-end text spotting results on the IC15 dataset. Note that, the F-measure with a generic lexicon (denoted as “G”) is the most important metric in the IC15 dataset. When the shorter side is 736 pixels, the speed of our method reaches 24.5 FPS, which is 3 to 7 times faster than previous methods. Meanwhile, it still has a competitive “G” F-measure of 66.2, surpassing the “G” F-measure of 69.2%, which is 1.2 points higher than the previous best text spotter [Qin et al. 16]. At the same time, our inference speed is still the fastest (13.8 FPS).

These results show that in addition to curved text, our PAN++ can also achieve an excellent balance between accuracy and speed when detecting and recognizing straight text. We present some qualitative straight text spotting results in Fig. 12(b).

4.4 Experiments on Chinese Text Dataset

4.4.1 Experiment Setting

We choose the RCTW-17 [61] dataset to test the effectiveness of our method in detecting and recognizing Chinese text, which is a commonly used language that is completely different from English. Due to the number of character categories is large (up to 3,000), the detection and recognition of Chinese text is challenging. RCTW-17 has the same detection metric as English text datasets (e.g., IC15 [25]), but it uses the average edit distance (AED) [61] as the recognition metric. For a fair comparison, we randomly selected 1,000 images from the RCTW-17 training set as the validation set, and trained all models for 300 epochs in the remaining training set without using external text datasets. Other model settings, data augmentations, and training settings are the same as those in Sec. 4.2. In the testing phase, we rescale the shorter side of the input image to 736 pixels.

4.4.2 Chinese Text Results

The text detection and recognition performance on the RCTW-17 dataset are reported in Table 5. With ResNet18 [19] as the backbone, our PAN++ achieves F-measure of 83.1% and 23.4 AED, which is 11.2 points (65.1% vs. 53.9%) and 1.7 points (23.4% vs. 25.1%) better than those of the baseline [SegLink 34].
TABLE 3
Text detection results on IC15 and MSRA-TD500. “Scale” denotes the scale of the testing image, where “L:” means the longer side is fixed, and “S:” means the shorter side is fixed, and “MS” means multi-scale testing. Note that, we use 1080Ti GPU to test the inference speed of PAN++ (detection only) for fair comparisons.

| Method          | Scale | Backbone | External Dataset | IC15  | MSRA-TD500 |
|-----------------|-------|----------|------------------|-------|------------|
|                 |       |          |                  | Precision | Recall | F-measure | FPS | Precision | Recall | F-measure | FPS |
| CTPN [12]       | S: 600 | VGG16    | ×                | 74.2   | 51.6 | 60.9 | 7.1 | - | - | - | - |
| RRPN [68]       | L: 1000 | VGG16   | ×                | 82.0   | 73.0 | 77.0 | - | 82.0 | 68.0 | 74.0 | - | - |
| EAST [14]       | 720×1280 | VGG16 | ×                | 83.6   | 73.5 | 78.2 | 13.2 | 87.3 | 67.4 | 76.1 | - | - |
| PixelLink [30]  | 768×1280 | VGG16 | ×                | 82.9   | 81.7 | 82.3 | 7.3 | 81.1 | 73.0 | 76.8 | 3.0 | - |
| DeepReg [70]    | MS    | VGG16    | ×                | 82.0   | 80.0 | 81.0 | - | 77.0 | 70.0 | 74.0 | 1.1 | - |
| PAN++ (ours)    | S: 736 | ResNet18 | × | 85.5   | 77.2 | 81.2 | 28.2 | 81.6 | 80.3 | 80.9 | 32.5 | 22.6 |
|                 | S: 896 | ResNet18 | × | 86.7   | 78.4 | 82.3 | 19.2 | 82.4 | 82.1 | 82.2 | 22.6 | - |
| SegLink [34]    | 768×1280 | VGG16 | SynthText | 73.1 | 76.8 | 75.0 | - | 86.0 | 70.0 | 77.0 | 8.9 | - |
| SSTD [71]       | 704×704 | VGG16 | Private Data | 80.2 | 73.9 | 76.9 | 7.7 | - | - | - | - |
| RRD [72]        | 1024×1024 | VGG16 | SynthText | 85.6 | 79.0 | 82.2 | 6.5 | 87.0 | 73.0 | 79.0 | 10 | - |
| MCN [73]        | 512×512 | VGG16 | SynthText | 72.0 | 80.0 | 76.0 | - | 88.0 | 79.0 | 83.0 | - | - |
| TextSnake [37]  | 768×1280 | VGG16 | SynthText | 84.9 | 80.4 | 82.6 | 1.1 | 83.2 | 73.9 | 78.3 | 1.1 | - |
| LOMO [57]       | L: 1536 | ResNet50 | SynthText | 91.3 | 83.5 | 87.2 | - | - | - | - | - |
| SPCNet [11]     | S: 800 | ResNet18 | IC17-MLT | 88.7 | 85.8 | 87.2 | 4.6 | - | - | - | - |
| PAN++ (ours)    | S: 736 | ResNet18 | SynthText | 85.9 | 80.4 | 83.1 | 28.2 | 85.3 | 84.0 | 84.7 | 32.5 | - |
|                 | S: 896 | ResNet18 | SynthText | 88.7 | 80.7 | 84.5 | 19.2 | 89.6 | 86.3 | 87.9 | 22.6 | - |
|                 | S: 896 | ResNet50 | IC17-MLT | 91.4 | 83.9 | 87.5 | 12.6 | 91.4 | 85.6 | 88.4 | 14.3 | - |

Fig. 11. Qualitative text detection results of PAN++ on Total-Text [23], CTW1500 [24], IC15 [25] and MSRA-TD500 [26]. Our method works well on various complex natural scenes, for example, arbitrary shapes, multiple languages, extreme illumination.

At the same time, our method runs 3 times faster than the baseline (18.5 FPS vs. 5.9 FPS). When using a heavier backbone (i.e., ResNet50 [19]) and a larger input scale, the text detection F-measure and AED of our method further improve to 68.0% and 22.2%, 14.1 points (68.0% vs. 53.9%) and 2.9 points (22.2% vs. 25.1%) better than the baseline, while keeping a faster inference speed (10.7 FPS vs. 5.9 FPS). Some qualitative Chinese text detection and recognition results are presented in Fig. 12 (c). These results prove that our PAN++ can also works well in Chinese text detection and recognition.

4.5 Comparison with Conference Versions

The key points of PAN++ and its conference versions are summarized as follows:

- **PSENet [4]** (Text Detector): Kernel Representation, PSE;
- **PAN [17]** (Text Detector): FPEM, PA;
- **PAN++ (E2E Text Spotter):** Masked RoI, Recognition Head, FPEM (improved), PA (improved);
- **CRNN [5]**. At the same time, our method runs 3 times faster than the baseline (18.5 FPS vs. 5.9 FPS). When using a heavier backbone (i.e., ResNet50 [19]) and a larger input scale, the text detection F-measure and AED of our method further improve to 68.0% and 22.2%, 14.1 points (68.0% vs. 53.9%) and 2.9 points (22.2% vs. 25.1%) better than the baseline, while keeping a faster inference speed (10.7 FPS vs. 5.9 FPS). Some qualitative Chinese text detection and recognition results are presented in Fig. 12 (c). These results prove that our PAN++ can also works well in Chinese text detection and recognition.

As an extension of conference versions, PAN++ evolves from the original text detector to an end-to-end (E2E) arbitrarily-shaped text spotter. To this end, we rebuild the overall architecture of PAN++ and carefully integrate a feature extractor (i.e., Masked RoI) and a lightweight recognition head in it.

Besides that, when inheriting the kernel representation, FPEM, and PA from the conference versions, we make some improvements as follows: 1) We systematically compare our kernel representation with other mainstream representations, and prove that our representation is friendly to real-time applications. 2) We simplify the FPEM by combining the old FPEM [17] and FFM [17] into a single module, which is more effective. 3) We improve PA and make it aware of background elements by adding a background item to the discrimination loss.

With the above-mentioned improvements, PAN++ can simultaneously detect and recognize arbitrarily-shaped text, which is impossible with models in conference versions. Moreover, PAN++ also achieve better text detection performance than its predeces-
End-to-end text spotting results on the IC15 dataset. “S” (strong) means 100 words, including the ground truth, are given for each image. For “W” (weak), a lexicon includes all words that appear in the test set is provided. For “G” (generic), a generic lexicon with 90k words is given. “N” (none) means that no lexicon is used. “Scale” denotes the scale of the testing image, where “L” means the longer side is fixed, and “S” means the shorter side is fixed, and “MS” means multi-scale testing.

| Method            | Scale | Backbone | Training Strategy | External Dataset | IC15 |
|-------------------|-------|----------|-------------------|------------------|------|
| He et al. [15]    | L: 2240 | ResNet50 | Finetune          | SynthText        | 81.1 | 75.9 | 60.8 | - | 7.5 |
| CharNet [31]      | L: 2280 | ResNet50 | Finetune          | SynthText, IC17-MLT, Total-Text | 82.8 | 78.9 | 67.6 | 62.7 | 1.2 |
| TextNet [69]      | MS    | ResNet50 | Finetext          | SynthText        | 78.7 | 74.9 | 60.5 | - | - |
| TextDragon [32]   | VGG16 | Finetune | SynthText, Total-Text | SynthText, Total-Text | 85.1 | 81.5 | 65.2 | - | - |
| PAN++ (ours)      | S: 756 | ResNet18 | Finetune          | SynthText, COCO-Text, IC17-MLT | 79.4 | 74.9 | 65.6 | 64.7 | 24.5 |
|                   | S: 896 | ResNet50 | Finetune          | SynthText, COCO-Text, IC17-MLT | 80.6 | 76.0 | 66.4 | 65.7 | 18.6 |
| Deep TextSpotter v2 [28] | 608 x 608 Darknet19 | Jointly Train | SynthText, IC13 | SynthText, IC13 | 54.0 | 51.0 | 47.0 | - | 9.0 |
| Mask TextSpotter     | S: 1000 | ResNet50 | Jointly Train | SynthText, IC13, Total-Text | 77.3 | 69.9 | 60.3 | - | 4.8 |
| Mask TextSpotter v2 [30] | S: 720 | ResNet50 | Jointly Train | SynthText, IC13, Total-Text, SCUT | 74.2 | 69.2 | 63.5 | - | 3.8 |
| Qin et al. [16]   | S: 900 | ResNet50 | Jointly Train | SynthText, COCO-Text, IC17-MLT, Total-Text, Private Data | 83.4 | 79.9 | 68.0 | - | 4.8 |
| PAN++ (ours)      | S: 756 | ResNet18 | Jointly Train | SynthText, COCO-Text, IC17-MLT, Total-Text | 80.5 | 75.8 | 66.2 | 65.7 | 24.5 |
|                   | S: 896 | ResNet18 | Jointly Train | SynthText, COCO-Text, IC17-MLT, Total-Text | 81.3 | 76.4 | 67.4 | 66.6 | 18.6 |
|                   | S: 896 | ResNet50 | Jointly Train | SynthText, COCO-Text, IC17-MLT, Total-Text | 82.7 | 78.2 | 69.2 | 68.0 | 13.8 |

TABLE 5
Results on the RCTW-17 dataset. “Scale” denotes the scale of the testing image, where “S” means the shorter side is fixed, “F” denotes the F-measure, “AED” represents average edit distance defined in [61].

| Method        | Scale | Backbone | RCTW-17 |
|---------------|-------|----------|---------|
| SegLink [34] + CRNN [5] | 768 x 1280 | VGG16 | 83.9 | 25.1 | 5.9 |
| PAN++ (ours) | S: 756 | ResNet18 | 65.4 | 25.4 | 18.5 |
|               | S: 896 | ResNet50 | 68.0 | 22.2 | 10.7 |

TABLE 6
Detection performance comparison of PAN++ and its conference versions. “S” means the shorter side is fixed. “F” denotes the F-measure. “impr” means the improved module in this work. At a similar or faster inference speed, the F-measure of PAN++ is significantly better than previous versions. Note that, the inference speed is tested on one 1080Ti GPU for fair comparisons.

| Method            | Total-Text (S: 640) | IC15 (S: 736) |
|-------------------|---------------------|---------------|
|                  | F | FPS | F | FPS |
| PSENet-1s [4]    | 78.1 | 4.5 | 76.9 | 3.8 |
| PAN [17]         | 83.5 | 39.6 | 80.3 | 26.1 |
| PAN + PA (impr)  | 84.0 | 38.2 | 80.7 | 28.0 |
| PAN++ (detection only): PAN + PA (impr) + FPEM (impr) | 84.5 | 38.3 | 81.2 | 28.2 |

4.6 Ablation Study

4.6.1 Experiment Settings
To analyze PAN++ in depth, we conduct ablation studies on both curved and straight text datasets (e.g., Total-Text and IC15). In...
4.6.2 Can text kernels be used as detection results?
As mentioned in Sec. 3.1, the text kernel is the central region of a text line, which is used to locate and separate adjacent text lines during the segmentation process. It cannot cover the complete shape of a text line and cannot be directly used as the detection result. As shown in Fig. 14, the text detection F-measure of the model that only uses text kernel is much lower than the models with the default setting. In addition, we use a text recognizer CRNN [5] to recognize the text in the text kernel, and we find that CRNN fails to recognize it as shown in Fig. 13.

These results demonstrate that the text kernel cannot be used alone as the detection result, and it needs to be combined with other parts (i.e., text regions and instance vectors) of the kernel representation.

4.6.3 Influence of the Shrinking Rate
To study the impact of the shrinking rate on the text detection performance, we compare the text detection F-measure of models with different shrinking rates. Specifically, we vary the shrinking rate $r$ from 0.1 to 1 and evaluate PAN++ (detection only) on Total-Text and IC15 datasets. As shown in Fig. 13, the F-measures on both datasets drop when the shrinking rate $r$ is too large or too small. When $r$ is too large (i.e., $r \geq 0.9$), text kernels may fail to separate the text lines lying close to each other. When $r$ is too small (i.e., $r \leq 0.2$), a whole text line may be incorrectly divided into several connected components in the testing phase. Experimentally, the shrinking rate is set to 0.5 on IC15, and 0.7 on other datasets by default.

Note that, when the shrinking rate $r$ is equal to 1, the text kernel is the same as the text region, and the detection module will degenerate to a binary segmentation network without PA. In this case, the text detection F-measure drops sharply on Total-Text and IC15, because the model loses the capability to distinguish adjacent text lines as shown in Fig. 1(b).

### Table 7

| $N_{stk}$ | Param. (M) | Total-Text | IC15 |
|-----------|------------|------------|------|
|           | Total-Text | IC15       |      |
| 0         | 13.1       | 64.5       | 0.6  | 63.7 | 0.7  |
| 1         | 13.2       | 66.0       | 2.2  | 64.2 | 2.8  |
| 2         | 13.3       | 66.4       | 3.3  | 64.7 | 4.4  |
| 3         | 13.4       | 66.5       | 5.1  | 64.8 | 6.6  |
| 4         | 13.5       | 66.6       | 6.5  | 64.9 | 8.7  |

4.6.4 Influence of the Stack Number of FPEMs
We study the effect of the stack number of FPEMs on text spotting performance, by varying the stack number $N_{stk}$ from 0 to 4. Note that, when the stack number $N_{stk}$ is equal to 0, there is no FPEM and the final feature map $F_f$ is generated by upsampling and concatenating the feature pyramid $F_r$ (see Fig. 4(c)). From Table 7 we find that the text spotting F-measures on both datasets keep rising with the growth of $N_{stk}$ and begins to level off when the stack number $N_{stk}$ is greater than 2. The stack number $N_{stk}$ is a trade-off between accuracy and speed. Although FPEM has a low computational cost, the model will slow down when the stack number $N_{stk}$ is large. Specifically, an additional FPEM takes up to 2ms per image. To keep a good balance between performance and speed, $N_{stk}$ is set to 2 by default.

4.6.5 Effectiveness of FPEM
To verify the effectiveness of FPEM, we first compare the performance of models with and without this module. As reported in Table 7, the model with 2-stacked FPEMs ($N_{stk} = 2$) obtains the F-measure of 66.4% on the Total-Text dataset, which is 1.9 points higher than that of the model without FPEM ($N_{stk} = 0$).

Second, we compare our feature enhancement network (2-stacked FPEMs) with other widely used methods, including FPN [51], PAFPN [74], and NAS-FPN [75]. For fair comparisons, we directly replace the feature enhancement network with other methods and evaluate them under the same experiment setting mentioned in Sec. 4.6.1. In Table 8 we find that our method achieves comparable or even better text spotting performance than counterparts, while our time cost and parameter number are much

Fig. 13. Recognition results of the complete bounding box and the text kernel of the same text line. CRNN [5] successfully recognizes the text in the complete bounding box but fail to recognize the text in the text kernel.

Fig. 14. Text detection F-measure of PAN++ under different shrink rates. When we raise the shrinking rate, the F-measure first increases and then decreases. In addition, the F-measure drops sharply when we only use text kernels as the detection result.
lower. Concretely, our 2-stacked FPEMs yields the end-to-end text spotting F-measure of 66.4% on Total-Text, which is better than FPN [51] (65.9%) and close to NAS-FPN [73] (66.9%). However, the time cost of our method is only 1/3 of FPN and 1/10 of NAS-FPN (3.8 ms vs. 12.1 ms and 41.2 ms), and our parameters is 1/2 of FPN and 1/4 of NAS-FPN (13.3 M vs. 20.3 M and 46.9 M).

In conclusion, the proposed FPEM can obtain a good balance between accuracy and inference speed/parameter number, benefiting from the separable convolution and stackable design. It is a good choice to enhance the feature from a lightweight backbone, at least for text detection and end-to-end text spotting tasks.

4.6.6 Effectiveness of PA

We study the effectiveness of PA by removing it from the pipeline. In detail, we set $\beta$ in Eqn. (13) to 0 during training and merge all neighbor text pixels in step 2 of post-processing. Comparing the model with PA (see Table 9 #1), the F-measure of the model without PA (see Table 9 #2) drops over 1.0 points, which indicates the effectiveness of PA.

In Fig. 15, we also present some text kernel results predicted by models with and without PA. The model with PA tends to predict better text kernels, reducing the incorrect cases shown in red circles.

4.6.7 Effectiveness of Masked RoI

As shown in Table 11 #1 and #2, we show the effectiveness of Masked RoI. Higher text spotting F-measure is achieved in the presence of Masked RoI. The model with the proposed Masked RoI outperforms the one with the naive RoI extractor (removing the masking operation in Masked RoI) by at least 1.1%. The underlying reason is that without Masked RoI, the model may fail to recognize text lines in a noisy background.

4.6.8 Effectiveness of Recognition Head

To test the effectiveness of our recognition head, we compare it with some state-of-the-art methods, including SAR [13] and Transformer [21]. For fair comparisons, we replace our recognition head with other methods and evaluate them under the same experiment setting mentioned in Sec. 4.6.1. As shown in Table 11 our recognition head outperform SAR [13] in terms of accuracy and inference speed. Especially in terms of speed, the time cost of our method is only 1/2 of SAR’s (2.3ms vs. 5.0ms), which is important when encountering the image with dense text. Compared with Transformer [21], our method yields a comparable F-measure (66.4% vs. 67.0%) at a speed of 2.3ms per text line, which is 10 times faster.

These results demonstrate that our recognition head keeps a good balance between accuracy and inference speed. It may benefit from the designs as follows: 1) Unlike SAR and Transformer, our method has no encoder, skipping the time cost of the encoder; 2) We adopt the multi-head attention layer, whose computational cost is lower than that of SAR; 3) Our decoder only contains two LSTM layers and one attention layer.

4.6.9 Influence of the Backbone Network

To analyze the capability of the proposed PAN++, we replace the lightweight backbone to heavier backbone (i.e., ResNet50 and VGG16). As shown in Table 9 #1, #3, and #4, heavier backbones can bring over 1 points improvement on both Total-Text and IC15 datasets. However, the reduction of FPS brought by the heavy backbone is also apparent.
and test on other datasets. Because this experiment aims to test and evaluate the model without recognition head trained on one dataset to further demonstrate the robustness of the proposed PAN++, we can focus on the correct position even when recognizing each step, which provides a great tool to analyze the performance of the recognition head. We find that the attention masks of the decoder are not accurate in locating the start of the string.

4.8 Robustness Analysis

To further demonstrate the robustness of the proposed PAN++, we evaluate the model without a recognition head trained on one dataset and test on other datasets. Because this experiment aims to test the generalization of our method, we do not use external datasets during training, and we also remove the recognition head that needs to be trained on the large-scale dataset. In the testing phase, we set the shorter sides of test images in Total-Text, CTW1500, IC15, and MSRA-TD500 datasets to 640, 640, 736, and 736, respectively.

Based on the annotation level, we divide the datasets into two groups, which are word level and text line level datasets. SynthText, ICDAR 2015, and Total-Text are annotated at the word level, and CTW1500 and MSRA-TD500 are annotated at the text line level.

The cross-dataset results of PAN++ are shown in Table 12. Notably, the model trained on SynthText (a synthetic dataset) has fair performance on IC15 and Total-Text, which indicates that even without any annotated data, our method can satisfy the scene text detection with low precision requirements. The model trained on manually annotated datasets has over 62 text detection F-measure in the cross-dataset evaluation, which is competitive. Furthermore, in the cross-dataset evaluation at the text line level, the F-measure of all models exceeds 78, even though these models are trained on curved/straight text datasets and tested on another dataset. These results demonstrate that the detection part of PAN++ can be well generalized to unseen datasets.

4.9 Speed Analysis

Table 13 present the time cost of all components of PAN++. We find that the time cost of text recognition accounts for half of the total time cost. Therefore, an obvious way to increase speed in practical applications is to run the recognition component in parallel through a basic producer-consumer model, which can reduce the time cost to the original 1/2. We conduct the speed analysis on Total-Text. We evaluate all testing images and calculate the average speed. These results are tested with 1 batch size on one V100 GPU and one 2.20GHz CPU in a single thread.

4.10 Failure Cases and Discussion

As demonstrated in previous experiments, the proposed PAN++ works well in most cases of arbitrarily-shaped text detection and recognition. It still fails for some difficult images, such as...
detecting large character spacing (see Fig. 17 (a)) or text-like areas (see Fig. 17 (b)) and recognizing texts in fancy font styles (see Fig. 17 (c)). Large character spacing is a problem widely existing in other state-of-the-art methods, which needs linguistic features to solve it. For text-like areas and text in fancy font styles, the main reason for them is the lack of training samples, and we believe that these problems will alleviate when we have sufficient training samples of them.

5 CONCLUSION

In this work, we have considerably extended the kernel representation for text spotting, which first appeared in our conference versions. We have also thoroughly discussed its feasibility in real-time arbitrarily-shaped text recognition. In particular, To further strengthen this representation, here we have developed an efficient framework for end-to-end arbitrarily-shaped text spotting, where we have managed to speed up all parts in the text spotting procedure by carefully designing a series of lightweight modules, including 1) a feature enhancement network with stacked FPEMs, 2) a detection head with PA, and 3) a recognition head with Masked RoI.

We have verified the performance of PAN++ on both text detection and end-to-end text spotting tasks. The experiment results demonstrate that our method offers advantages in both accuracy and inference speed compared with a few previous state-of-the-art methods. We hope that our PAN++ can serve as a cornerstone for real-world text understanding applications.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China under Grant 61672273 and Grant 61832008, the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant BK20160021. This work was in part supported by Alibaba Group through Alibaba Innovative Research (AIR) Program.

REFERENCES

[1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[2] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comp. Vis., 2016.
[3] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “EAST: an efficient and accurate scene text detector,” in Proc. IEEE Conf. Comp. Vis. Patt. Recog., 2017.
[4] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, “Shape robust text detection with progressive scale expansion network,” in Proc. IEEE Conf. Comp. Vis. Patt. Recog., 2019.
[5] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 11, pp. 2298–2304, 2016.
[6] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, “Aster: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2035–2048, 2018.
[7] 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. Comp. Vis. Patt. Recog., 2018.
[8] 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. Comp. Vis., 2018.
[9] M. Liao, B. Shi, and X. Bai, “Textboxes++: A single-shot oriented scene text detector,” IEEE Trans. Image Process., vol. 27, no. 8, pp. 3676–3690, 2018.
[10] J. Long, E. Shelhamer, and T. Darrell. “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comp. Vis. Patt. Recog., 2015.
[11] E. Xie, Y. Zang, S. Shao, G. Yu, C. Yao, and G. Li, “Scene text detection with supervised pyramid context network,” in Proc. AAAI Conf. Artificial Intell., 2019.
[12] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network,” in Proc. Eur. Conf. Comp. Vis., 2016.
[13] H. Li, P. Wang, C. Shen, and G. Zhang, “Show, attend and read: A simple and strong baseline for irregular text recognition,” in Proc. AAAI Conf. Artificial Intell., 2019.
[68] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary-oriented scene text detection via rotation proposals,” *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[69] Y. Sun, C. Zhang, Z. Huang, J. Liu, J. Han, and E. Ding, “Textnet: Irregular text reading from images with an end-to-end trainable network,” in *Proc. Asian Conf. Comp. Vis.*, 2018.

[70] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, “Deep direct regression for multi-oriented scene text detection,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017.

[71] 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. Comp. Vis.*, 2017.

[72] M. Liao, Z. Zhu, B. Shi, G.-s. Xia, and X. Bai, “Rotation-sensitive regression for oriented scene text detection,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2018.

[73] Z. Liu, G. Lin, S. Yang, J. Feng, W. Lin, and W. L. Goh, “Learning markov clustering networks for scene text detection,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2018.

[74] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2018.

[75] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2019.