This ACCV 2020 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2020
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Scale-Aware Polar Representation for Arbitrarily-Shaped Text Detection

Yanguang Bi and Zhiqiang Hu✉

SenseTime Research
{biyanguang,huzhiqiang}@sensetime.com

Abstract. Arbitrarily-shaped text detection faces two major challenges: 1) various scales and 2) irregular angles. Previous works regress the text boundary in Cartesian coordinates as ordinary object detection. However, such grid space interleaves the unique scale and angle attributes of text, which seriously affects detection performance. The implicit disregard of text scale also impairs multi-scale detection ability. To better learn the arbitrary text boundary and handle the text scale variation, we propose a novel Scale-Aware Polar Representation (SAPR) framework. The text boundary is represented in Polar coordinates, where scale and angle of text could be both clearly expressed for targeted learning. This simple but effective transformation brings significant performance improvement. The explicit learning on separated text scale also promotes the multi-scale detection ability. Based on the Polar representation, we design line IoU loss and symmetry sine loss to better optimize the scale and angle of text with a multi-path decoder architecture. Furthermore, an accurate center line calculation is proposed to guide text boundary restoration under various scales. Overall, the proposed SAPR framework is able to effectively detect arbitrarily-shaped texts and tackle the scale variation simultaneously. The state-of-the-art results on multiple benchmarks solidly demonstrate the effectiveness and superiority of SAPR.

1 Introduction

Scene text detection plays an important role in numerous applications, such as real-time text translation, product identification, image retrieve and autonomous driving. In recent years, deep learning based methods exhibit promising detection performance [1–4]. The promotion mainly benefits from the development of Convolutional Neural Networks (CNN) [5] and research on detection [6–8] and segmentation [9, 10]. However, many existing methods with quadrilateral bounding-box outputs may suffer from texts with arbitrary shapes. Consequently, more and more recent works [11–18] begin to focus on arbitrarily-shaped text detection.

The two basic and independent attributes of arbitrarily-shaped text are: 1) various scales and 2) irregular angles, which also become the major challenges in detection task. As shown in Fig. 1, the text prototypes could form arbitrary shapes based on scale and angle transformations. Since text is usually composed of multiple individual characters, the formulation of arbitrary boundary in Fig. 1

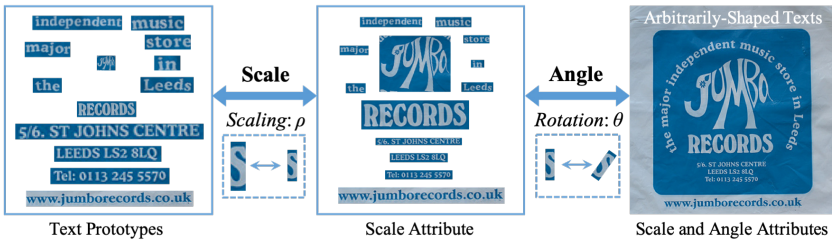

Fig. 1. Scale and angle attributes are the two basic and independent attributes of text boundary, which could form arbitrary shapes. From a decomposition perspective, the shape formulation is actually the scaling and rotation of small local areas.

is actually the simple scaling and rotation of small local areas, which could be naturally expressed in Polar coordinates. Therefore, we are motivated to detect arbitrarily-shaped text using Polar representation.

Fig. 2 presents the comparison of existing regular Cartesian representation and our novel Polar representation. For the scene text in Fig. 2(a), many methods [19, 20, 15, 21] regress the text boundary in Cartesian coordinates, as shown in Fig. 2(b)-(c). Compared with Fig. 2(b), Fig. 2(c) is more reasonable and stable which only focuses on the nearest boundary. While this is common in ordinary object detection, the unique scale and angle attributes of text are interleaved in such grid space. The force learning of unrelated attributes seriously affects detection performance, as shown in following experiments. The scale attribute of text is also implicitly disregarded, which impairs multi-scale detection ability. The overall detection performance in Cartesian space is thus largely suppressed. Conversely, Fig. 2(d) shows our Polar representation for arbitrarily-shaped texts. The independent scale and angle attributes are both clearly expressed, which is beneficial to boundary learning. Furthermore, the scale attribute is explicitly extracted and allows more effective end-to-end optimization. The multi-scale detection ability is thus promoted. On the whole, the transformation from Cartesian space to Polar space is simple but effective, which breaks the grid bottlenecks and brings significant performance improvement.

Based on the novel Polar representation, a unified Scale-Aware Polar Representation (SAPR) framework is proposed to better detect arbitrarily-shaped text and handle scale variation simultaneously. We dedicatedly design line IoU loss and symmetry sine loss to optimize the scale and angle attributes of text, respectively. Compared with L_1 loss and monotonous cosine loss, the tailored losses bring more performance improvement. A novel network architecture with multi-path decoder is also developed to better extract features from different scales. Besides, we propose a more accurate calculation of text center line which is frequently used in text detection task to complete entire boundary. Instead of complicated network prediction, we simply encode the symmetry distances of scale attribute. The produced center line could automatically fit various scales. Integrating above work, the unified SAPR framework is able to effectively detect

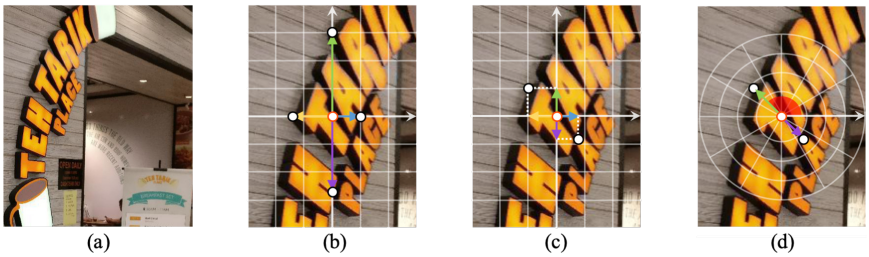

Fig. 2. (a) Arbitrarily-shaped texts. (b)-(c) Cartesian representations on global and local text. (d) The proposed Polar representation, which clearly depicts angle and scale attributes of text. Specifically, the scale attribute in (d) is decomposed into top distance (green arrow) and bottom distance (purple arrow).

texts with arbitrary shapes and handle the scale variation. The state-of-the-art empirical results on different benchmarks, especially the large improvement on arbitrarily-shaped datasets, demonstrate the effectiveness of SAPR.

The contributions of this work are summarized as follows: (1) We propose a novel Polar representation to better model arbitrary text boundary and learn the scale attribute simultaneously; (2) Based on the Polar representation, we develop line IoU loss and symmetry sine loss with multi-path decoder architecture as a unified Scale-Aware Polar Representation (SAPR) framework; (3) Instead of learning segmentation or attractive links, we proposed a more accurate and simple text center line extraction based on the symmetry distances of scale attribute; (4) SAPR achieves state-of-the-art performances on challenging Total-Text, CTW1500 and MSRA-TD500 benchmarks, which contain curved, multi-oriented and long texts.

2 Related Work

In recent years, most of the scene text detection methods are based on deep learning. They can be roughly divided into two categories: regression based methods and segmentation based methods.

Regression based methods benefit from the development of general object detection. Inspired by the Faster RCNN [7], CTPN [1] detects horizontal texts by grouping adjacent and compact text components. TextBoxes [22] and RRD [23] adopt the architecture of SSD [8] to detect texts with different aspect ratios. As the anchor-free methods, EAST [2] and DeepReg [24] predict the text boundary directly, which is similar to DenseBox [6]. RRPN [3] generates proposals with different rotations to detect multi-oriented texts. PMTD [4] is built on Mask RCNN [25] and produces quadrilateral boundary from pyramid mask. [26] detects texts by localizing corner points of bounding boxes. Most of the regression-based methods only predict quadrilateral bounding boxes with fixed number of vertexes. Therefore, such methods are difficult to detect texts

with arbitrary shapes. Besides, the limited receptive field of CNN also affects the detection performance on long texts.

Segmentation based methods benefit from the development of semantic segmentation. The Fully Convolutional Network (FCN) [9] and U-Net [10] are widely used structures. These methods aim to segment the text instances from backgrounds. For example, PixelLink [27] predicts the pixel classification and its neighborhood connections to obtain instances. With the rise of arbitrarily-shaped text detection trend, segmentation based methods become the mainstream because pixel-level classification is friendly to irregular shapes. However, the segmentation may cause adhesion when two text instances are close. Therefore, most of the segmentation based methods struggle to split adjacent texts. TextSnake [11], MSR [20] and LOMO [15] segment the center region and restore the boundary based on their regression results. TextField [17] predicts directional field to aggravate different instances. PSENet [28] predicts multiple kernels with different sizes and gradually merge them to produce final result. TextMountain [18] segment the center region of texts which are unconnected, then assign boundary pixels to corresponding center.

It is worth noting that compared with heuristic TextSnake, our method learns the text boundary end-to-end using polar representation. TextSnake and LOMO limit the shape regression to center region, while our method adaptively represents arbitrary texts anywhere. Moreover, our center line is calculated automatically with the symmetry scale distances in polar representation, which avoids extra complicated center learning.

3 Method

In this section, we first introduce the entire pipeline of SAPR framework. Next, the structure of network with multi-path decoder and the loss functions tailored for Polar representation are introduced. Then, the reconstruction of complete text boundary is presented in details. Finally, the label generation is described.

3.1 Scale-Aware Polar Representation Framework

The entire pipeline of SAPR framework is presented in Fig. 3. Overall, SAPR employs classification branch as mask to roughly locate texts and employs regression branch to precisely refine the boundary in Polar space. Specially, the scale attribute is decomposed into top distance and bottom distance, as shown by the green and purple arrows in Fig. 2(d). The angle is defined as the counter-clockwise rotation along the positive half axis. For an input image, the network produces text confidence from classification branch and angle, top distance, bottom distance from regression branch. The text confidence map is segmented to obtain text mask, which is used to cover valid text regions in regression maps. Based on top distance and bottom distance, we calculate the center line and extract the skeleton. Each individual center line skeleton is used to integrate local boundary restored by regression results and form a complete text boundary.

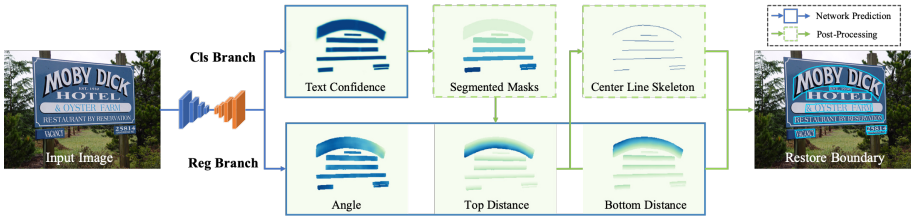

Fig. 3. The overview of SAPR framework. The center line skeleton is obtained automatically based on top distance and bottom distance, which guides the entire boundary restoration. The blue boxes with solid lines denote the outputs of network. The green boxes with dash lines denote post-processing.

Based on the suitable representation in Polar space, SAPR could better learn arbitrary text boundary and handle the scale problem compared with Cartesian methods [19, 20, 15, 21]. Different from heuristic approach [11], SAPR directly learns the text boundary with more effective and end-to-end manner which also simplifies the boundary restoration. In addition, the center text line is calculated easily and accurately with symmetry distances of scale attribute, which avoids complicated network learning of segmentation [11, 15] or attractive links [29, 30]. It is worth noting that the instance segmentation [31] also employ the polar representation with single center and fixed angle prior. However, the above simple representation is not suitable to curved texts with complex ribbon shapes. In contrast, our polar representation with various local centers and flexible angles could precisely describe irregular boundary and obtain promising performance.

On the whole, the Polar representation artfully express the scale and angle attributes of arbitrarily-shaped texts. Many bottlenecks in ordinary Cartesian space are solved gracefully. Therefore, SAPR achieves significant improvement of detection performance.

3.2 Network

The network of SAPR follows the typical encoder-decoder structure shown in Fig. 4. As a powerful feature extractor, the encoder produces rich feature maps with multiple levels. Generally, single path structure like U-Net is employed as the decoder. However, simple decoder may be too weak to process the rich and abstract information passed from encoder under complex multi-task learning. Besides, the high-level semantic information from the deep layers would also be diluted gradually during fusion.

Inspired by DLA [32] and GridNet [33], we develop a new decoder with multiple paths to better utilize information under different scales. During decoding, each path creates new aggregated features which are passed to next path via residual connections. We concatenate outputs from different paths for two parallel branches: text/non-text classification and shape regression. The multi-path decoder has more powerful representation ability to extract and analyze information from encoder for abstract and complex regression. At the same time,

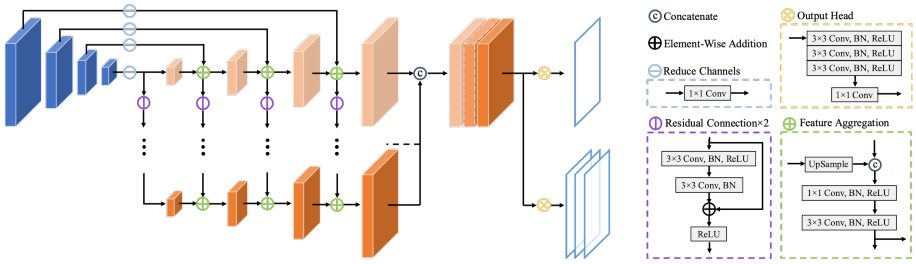

Fig. 4. The detailed structure of network with multi-path decoder. "Conv", "BN", "ReLU" and "UpSample" denote convolution, batch normalization, rectified linear unit and 2× bilinear up-sampling, respectively.

residual connections allow network to automatically learn the utilization degree of features in different scales. Thus the aggregation produces more effective features for multi-scale text detection.

3.3 Loss Function

The loss for multi-task learning is formulated as

$$L = \lambda_1 L_{cls} + \lambda_2 L_{dis} + \lambda_3 L_\theta \tag{1}$$

where L_{cls}, L_{dis} and L_θ represent the losses of text/non-text classification, distance regression and angle regression, respectively.

Binary cross-entropy loss shown in Eq. 2 is employed as the classification loss for fair comparison. $\mathcal{M}$ is the training mask to ignore invalid regions. $\widehat{y_i}$ and y_i denote ground truth and predicted label in the ith location, respectively.

$$L_{cls} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left(-\widehat{y_i} \log y_i - (1 - \widehat{y_i}) \log (1 - y_i)\right) \tag{2}$$

Instead of using norm losses like L_2, L_1 and *Smooth* L_1, we design line IoU loss, i.e., the one-dimensional (1D) version of original IoU loss [34] for top and bottom distances regression. In Eq. 5, the $\widehat{d}_i^{top}$, $\widehat{d}_i^{bot}$, d_i^{top} and d_i^{bot} denote top distance label, bottom distance label, predicted top distance and predicted bottom distance in the ith location, respectively. $\mathcal{T}$ denotes valid text regions. The proposed line IoU loss could better handle texts with various heights and thus contributes to the multi-scale detection.

$$d_i^{inter} = \min\left(\widehat{d}_i^{top}, d_i^{top}\right) + \min\left(\widehat{d}_i^{bot}, d_i^{bot}\right) \tag{3}$$

$$\widehat{d}_i = \widehat{d}_i^{top} + \widehat{d}_i^{bot}, d_i = d_i^{top} + d_i^{bot} \tag{4}$$

$$L_{dis} = -\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \log\left(\frac{d_i^{inter}}{\widehat{d}_i + d_i - d_i^{inter}}\right) \tag{5}$$

In horizontal text areas, the angle may change dramatically between 180° and 0°. The angle label is thus discontinuous. However, the actual shape appearances of texts in such areas are stable. It is both reasonable to predict θ or $\pi - \theta$ in these areas. Therefore, we design symmetry sine loss in Eq. 6 to alleviate the confusion in transition areas and make network easier to converge. $\widehat{\theta_i}$ and θ_i denote the angle label and predicted angle in the ith location, respectively.

$$L_\theta = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \sin(|\widehat{\theta_i} - \theta_i|) \tag{6}$$

In the whole training stage, λ_1, λ_2 and λ_3 are set to 10, 1 and 1, respectively.

3.4 Label Generation

The labels includes: (1) text confidence, (2) top distance, (3) bottom distance and (4) angle. Fig. 5 shows the detailed label generation from polygon annotation.

Assuming that the annotations only include ordered vertexes. The preparation of label generation is finding the 4 key vertexes denoted by the green numbers 1-4 in Fig. 5. For convenience, each text instance annotation is formulated as $\mathcal{V} = \{v_1, ..., v_i, ..., v_n\}$ and $n \geq 4$, where n denotes the number of polygon vertexes. Each vertex v_i can be viewed as the intersection point of two adjacent sides denoted by $\vec{v_l}$ and $\vec{v_r}$. We define θ_i in Eq. 7 to measure degree of direction change of these two adjacent sides. The v_i is more likely to being a key vertex when θ_i is smaller, i.e., the adjacent sides construct a 90° angle.

$$\theta_i = |90° - \arccos\left(\vec{v_l} \cdot \vec{v_r} / \left(\|\vec{v_l}\| \, \|\vec{v_r}\|\right)\right)| \tag{7}$$

The entire label generation requires the nearest two side points p_t and p_b on polygon annotation in the normal direction, which are the white points in Fig. 5. Firstly, we construct the two paths between key vertexes 1-2 and key vertexes 3-4, denoted by l_t and l_b. Then, these two paths are sampled densely as $l_t = \{p_{t1}, ..., p_{ti}...p_{tm}\}$ and $l_b = \{p_{b1}, ..., p_{bi}...p_{bm}\}$ where m is the number of sampled points. For any location p in text region, we can calculate the distance d_i between p and the each line determined by (p_{ti}, p_{bi}). In this way, the nearest two side points p_t and p_b assigned for current location p could be obtained as:

$$p_t = p_{t\widehat{i}}, p_b = p_{b\widehat{i}}, \tag{8}$$

$$\widehat{i} = \arg\min_i d_i. \tag{9}$$

The text confidence label is the height-shrunk version of complete text mask, while the length on reading direction remains unchanged. For the height-shrunk mask, a appropriate ratio helps avoiding adhesion and ambiguous regression in the edge area compared with the original ratio=0 mask. A big ratio would cause confusion to classify the surrounding text as background. The ratio is not sensitive and works well in [0.15, 0.4], so we chose 0.3. The top distance label (green arrow) and bottom distance label (purple arrow) are the lengths between current location p with the higher side point p_t and the lower side point p_b, respectively. The angle label (red sector) is the angle between the local normal vector $\langle p_b, p_t \rangle$ and horizontal direction which ranges from 0° to 180°.

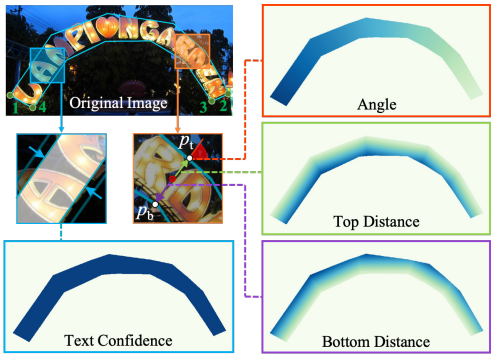

Fig. 5. The label generation based on original polygon annotation. The text confidence label is the height-shrunk version of complete text mask. The top distance and bottom distance labels are the lengths between current location p with the higher side point p_t and the lower side point p_b, respectively. The angle label is the angle between the local normal vector $\langle p_b, p_t \rangle$ with horizontal direction.

3.5 Text Boundary Restoration

Fig. 6 presents the detailed text boundary restoration. Firstly, the centerness is calculated based on top distance and bottom distance:

$$c_i = \frac{2 * \min\left(d_i^{top}, d_i^{bot}\right)}{d_i^{top} + d_i^{bot}} \tag{10}$$

where d_i^{top}, d_i^{bot} and c_i denote predicted top distance, predicted bottom distance and centerness in the ith location, respectively. The centerness ranges in [0, 1] with mountain appearance where the regions closer to center have larger values. Then, the center line could be easily segmented and skeletonized from centerness map. Each center line skeleton is considered as a individual instance to avoid adhesion. For example, the two small texts in the bottom right corner of Fig. 6 are dense and stick together. The skeleton successfully separate the adhesive texts. Next, the anchor points are sampled on each center line skeleton evenly. Based on the Polar coordinates regression in its surrounding region, each anchor point could directly produce corresponding local boundary.

Specifically, the transformation from predicted Polar coordinates to Cartesian coordinates contains two steps: 1) scale restoration and 2) angle restoration. Assuming that (x_i, y_i) denotes ith location in original image, where the regressed Polar coordinates are top distance d_i^{top}, bottom distance d_i^{bot} and angle θ_i. It is noteworthy that θ_i is the counterclockwise rotation along the positive half axis. The Cartesian coordinates of two local boundary points (x_i^t, y_i^t) and $\left(x_i^b, y_i^b\right)$ are:

$$\begin{bmatrix} x_i^t & x_i^b \\ y_i^t & y_i^b \end{bmatrix} = \begin{bmatrix} \cos\Delta\theta & -\sin\Delta\theta \\ \sin\Delta\theta & \cos\Delta\theta \end{bmatrix} \begin{bmatrix} 0 & 0 \\ d_i^{top} & -d_i^{bot} \end{bmatrix} + \begin{bmatrix} x_i \\ y_i \end{bmatrix} \tag{11}$$

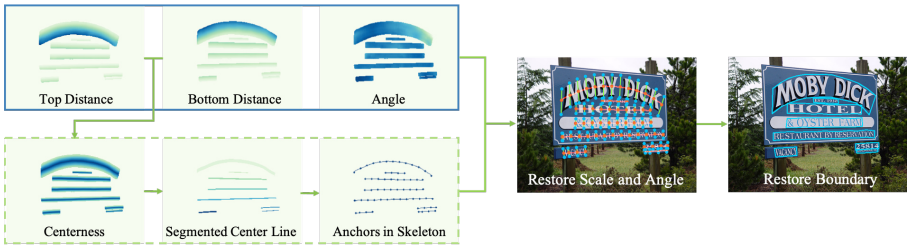

Fig. 6. The details of text boundary restoration. Based on top distance and bottom distance, centerness is calculated and segmented to automatically obtain center line. Anchor points are sampled on each skeleton line and restore scale and angle attributes. The complete boundary is formed by integrating local boundaries.

where $\Delta\theta = \theta - \frac{\pi}{2}$. The complete detection boundary is obtained by integrating the local boundary points.

Overall, the encoded centerness automatically produces accurate center line under different scales. The extracted skeletons naturally solve the adhesion problem. Besides, it avoids complicated network prediction of segmentation [11, 15] or attractive links [29, 30] to integrate local boundaries. The entire text boundary restoration is thus simplified and more robust.

4 Experiments

4.1 Datasets

SynthText [35] is a synthetic dataset which contains about 800K synthetic images. The texts are artificially rendered with random attributes. Like other methods, SynthText is used to pre-train our network.
Total-Text [36] is a recently released word-level dataset. It consists 1255 training images and 300 testing images with horizontal, multi-oriented, and curved texts. The annotations are polygons with variable vertexes.
CTW1500 [37] is a text-line based text dataset with 1000 training images and 500 testing images. Similar to Total-Text, the texts are also horizontal, multi-oriented, and curved. The annotations are polygons with fixed 14 vertexes.
MSRA-TD500 [38] is a line-level dataset with 300 training images and 200 test images of multi-oriented and long texts. The annotations are rotated rectangles. The training set is relatively small, so we also include 400 images from HUST-TR400 [39] according to previous works [2, 26, 11].

4.2 Implementation Details

We use ResNet50 [5] pre-trained on ImageNet [40] as the backbone of network, which produces 4 levels feature maps denoted by C_2, C_3, C_4 and C_5. Their channels are reduced to 32, 64, 128 and 256, respectively. The decoder is equipped

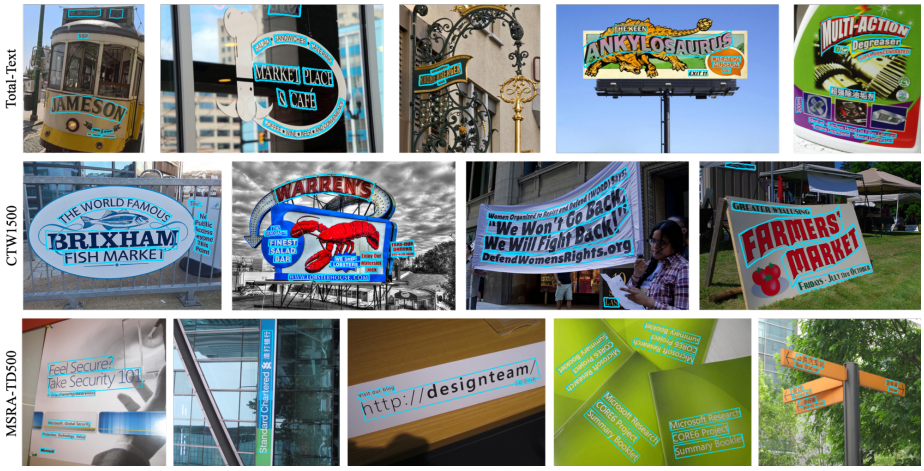

Fig. 7. Qualitative results of SAPR on different benchmarks. From top to bottom in rows: Total-Text, CTW1500, MSRA-TD500.

with 3 paths. The output channel of aggregation module is the same as the minimum channel number of two input features. The final outputs have strides of 4 pixels with respect to the input image.

The training contains two phase: pre-train and fine-tune. The model is optimized using ADAM with batch-size 32 and the learning rate decreases under cosine schedule. We use SynthText [35] to pre-train the model for 1 epoch. The learning rate is from 1×10^{-3} to 1×10^{-4}. Then, we fine-tune the model on different benchmarks for 100 epochs, respectively. The learning rate is from 3×10^{-4} to 1×10^{-5}. The blurred texts labeled as DO NOT CARE are ignored in training.

During the data augmentation, we set the short sides of training images in [640, 1280] randomly. The heights of images are set in ratio [0.7, 1.3] randomly while widths keep unchanged. The images are rotated in $[-15°, 15°]$ randomly. 640×640 random patches are cropped as the final training data.

During the evaluation on three benchmarks, the short sides of images are all fixed to 960 to report the single scale results. The evaluation on MSRA-TD500 requires box with fixed 4 vertexes, so we extract rotated rectangle with the minimum area around original detected polygon as the final result.

4.3 Comparisons with State-of-the-Art Methods

Arbitrarily-Shaped Text. Total-Text and CTW1500 are recently introduced datasets with arbitrarily-shaped texts. They are specially curated as the two most important benchmarks to evaluate the arbitrarily-shaped text detection performance. The quantitative comparisons on Total-Text and CTW1500 are shown in Table 1. SAPR achieves the new state-of-the-art performances on both challenging datasets with significant improvements. SAPR designs better polar

Table 1. Quantitative results of different methods on Total-Text, CTW1500 and MSRA-TD500 benchmarks. The best score is highlighted in bold.

Datasets	Total-Text			CTW1500			MSRA-TD500		
Method	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
SegLink [29]	30.3	23.8	26.7	-	-	-	86.0	70.0	77.0
EAST [2]	50.0	36.2	42.0	-	-	-	87.3	67.4	76.1
CENet [41]	59.9	54.4	57.0	-	-	-	85.9	75.3	80.2
CTD [42]	74.0	71.0	73.0	74.3	65.2	69.5	84.5	77.1	80.6
SLPR [19]	-	-	-	80.1	70.1	74.8	-	-	-
TextSnake [11]	82.7	74.5	78.4	67.9	**85.3**	75.6	83.2	73.9	78.3
SAE [13]	-	-	-	82.7	77.8	80.1	84.2	**81.7**	82.9
MSR [20]	85.2	73.0	78.6	84.1	79.0	81.5	87.4	76.7	81.7
AGBL [43]	84.9	73.5	78.8	83.9	76.6	80.1	86.6	72.0	78.6
TextField [17]	81.2	79.9	80.6	83.0	79.8	81.4	87.4	75.9	81.3
PSENet [28]	84.0	78.0	80.9	84.8	79.7	82.2	-	-	-
FTSN [44]	84.7	78.0	81.3	-	-	-	87.6	77.1	82.0
SWSL [45]	80.6	82.3	81.4	77.0	79.9	78.5	-	-	-
SegLink++ [30]	82.1	80.9	81.5	82.8	79.8	81.3	-	-	-
IncepText [46]	-	-	-	-	-	-	87.5	79.0	83.0
MCN [47]	-	-	-	-	-	-	88.0	79.0	83.0
Relation [48]	86.0	80.5	83.1	85.8	80.9	83.3	87.2	79.4	83.1
LOMO [15]	87.6	79.3	83.3	85.7	76.5	80.8	-	-	-
CRAFT [16]	87.6	79.9	83.6	86.0	81.1	83.5	88.2	78.2	82.9
ContourNet [49]	86.9	**83.9**	85.4	83.7	84.1	83.9	-	-	-
SAPR(ours)	**89.5**	82.6	**85.9**	**89.8**	83.2	**86.4**	**92.5**	75.7	**83.3**

representation thus output boundaries are more accurate with high confidence. Therefore, we use higher threshold to show precision superiority while suppress recall. Some detection results are presented in the first and second rows of Fig. 7, where arbitrarily-shaped texts under different scales are all precisely located. It solidly demonstrates the effectiveness of Polar representation and tailored framework on arbitrarily-shaped text detection.

Regular Text. We also evaluate SAPR on regular quadrilateral texts to prove the generalization ability. Among the different regular datasets, MSRA-TD500 is more challenging with a large amount of multi-oriented and extreme long texts. The quantitative comparison result is shown in Table 1 and SAPR still achieves the state-of-the-art performance. Some detection results are presented in the third row of Fig. 7, where the texts with multiple orientations and extreme long lengths are all accurately detected under various scales. This comparison indicates that SAPR could also seamlessly adapt to other types of text detection.

4.4 Ablation Study

Scale-Aware Polar Representation. To exhibit the effectiveness of SAPR, we train a similar model using the Cartesian representation [20, 15] in Fig. 2(c) as baseline while other configurations remain the same. Meanwhile, we divide

Fig. 8. The qualitative comparison between Cartesian baseline and SAPR under different scales. The red arrow, green arrow and yellow arrow denote false alarm, miss and inaccurate boundary, respectively.

Table 2. Comparison of Cartesian baseline and SAPR framework. "F", "F-small", "F-medium" and "F-large" denote the overall F-score, F-scores of small, medium and large texts, respectively.

Method	Polar Space	Multi-Path Decoder	Line IoU Loss	Symmetry Sine Loss	F	F-small	F-medium	F-large
Cartesian Baseline	-		-	-	77.3	66.1	74.1	78.4
	-	✓	-	-	78.4	68.5	75.4	79.6
SAPR(ours)	✓				81.6	70.4	81.1	84.0
	✓	✓			82.7	72.7	82.4	84.1
	✓	✓	✓		85.3	77.3	**85.3**	87.1
	✓	✓	✓	✓	**85.9**	**77.9**	84.5	**87.5**

texts into small, medium and large inspired by COCO dataset [50] to clearly demonstrate the scale-aware ability. For Total-Text, the texts with height in (0, 27), (27, 50) and (50,∞) are defined as small, medium and large, which occupy 42%, 31% and 27% of entire dataset, respectively.

Table 2 presents the detailed ablation results. During the evaluation, we use single path decoder, L_1 loss and monotonous cosine loss [2] as substitutes. The Cartesian baseline obtains 77.3% F-score. When our Polar representation is adopted, the F-score obviously increases to 81.6%. The F-scores of all three scale ranges also obtain promising improvement. It solidly demonstrates that the proposed Polar representation clearly decouple the independent scale and angle attributes of text, which is more suitable for learning distinguish features and improves detection performance. At the same time, the explicit learning of scale attribute also contributes to multi-scale detection ability.

Furthermore, the multi-path decoder promotes both the performances of Cartesian baseline and SAPR. It indicates the effectiveness of poposed multi-path structure to extract features from multiple scales. With the line IoU loss, the overall F-score increases obviously to 85.3% where the F-small obtains around

5% absolute improvement. Moreover, symmetry sine loss brings 0.6% absolute improvement. It proves that the tailored losses could effectively describe and optimize the scale and angle attributes of text. On the whole, the F-scores of overall and different scale ranges are gradually improved with proposed components. It demonstrates the effectiveness of SAPR framework to detect arbitrarily-shaped texts and handle the scale variation problem. In particular, the evaluation tool DetEval [51] allows many-to-one and one-to-many matches which may slightly affect the detection results in current scale range [52].

Fig. 8 shows qualitative comparison between Cartesian baseline and SAPR under different scales. The decoupling learning of independent angle and scale attributes is beneficial for network to explore more essential features of texts, which effectively reduces the false alarms (red arrow). Meanwhile, the miss of vague and indistinguishable text (green arrow) which are easy to mix with background are also improved. In addition, the Cartesian baseline may produce inaccurate boundaries with low quality (yellow arrow). By contrast, SAPR precisely locates texts with arbitrary shapes and various scales, which solidly demonstrates the superior performance of SAPR.

Table 3. The comparison of mean and variance of F-scores over dataset with multiple scale-fluctuations.

Method	F-score Mean	F-score Variance
Cartesian Baseline	75.1	3.17
SAPR(ours)	**83.7**	**2.51**

Although SAPR exhibits promising performances on different benchmarks, there are two types of fail cases: (1) The texts with blurry appearance are difficult to distinguish by the classification branch; (2) The boundary of too small/short texts produced by the regression branch may be not accurate and unstable.

Scale Robustness. To further confirm the scale-aware ability, the short sides of input images in Total-Text dataset are set in [640, 1280] with step 40 to obtain the F-score fluctuation as metric. Table 3 presents the mean and variance of F-scores with multiple scale-fluctuations. SAPR achieves 8.6% absolute F-score improvement compared with Cartesian baseline. Meanwhile, the variance also decreases obviously. It solidly demonstrates that the Polar representation clearly separates scale attribute of text, which allows more focused and effective learning on scale problem. With tailored line IoU loss, SAPR is able to be robust to complex scale variation.

Center Line Generation. We propose to use local top and bottom distance to encode the center line of text. For comparison, the classification branch is added a new channel to predict center line as baseline. During inference, the center line is directly segmented from the output of corresponding channel. As shown in Table 5, encoding the distance to produce center line achieves 5.4% absolute improvement compared with direct segmentation. Actually, it is hard

Table 4. Comparison of decoder with paths of different numbers.

#Path	1	2	3	4
Params(M)	35.44	34.43	32.53	33.59
Flops(G)	11.04	7.95	7.43	8.14
F-score	83.5	84.4	**85.9**	83.7

Table 5. Comparison of center line generation between common direct segmentation and our distance encoding.

Center Line	Precision	Recall	F-score
Direct Segmentation	83.9	77.3	80.5
Distance Encode	**89.5**	**82.6**	**85.9**

for network to learn the accurate center line, which is imagined artificially and prone to confused by the similar texts in surrounding sides. From the perspective of symmetry distance, encoding the top and bottom could naturally and easily produce reasonable center line with better quality.

Number of Decoder Path. The number of paths in decoder would affect final performance. We design decoders with paths of different numbers to find the best configuration. The parameters and flops (224×224 input) are tried to set similar by adjusting channels for fair comparison. Table 4 shows the comparison on Total-Text dataset. Compared with single-path decoder, multi-path decoders usually have better performance. However, more paths may be too complex to be trained efficiently with limited data. The decoder with 3 paths achieves the highest F-score, which is selected as our default configuration.

5 Conclusion

In this paper, we reveal the basic independent attributes of arbitrarily-shaped text boundary: 1) various scales and 2) irregular angles. Cartesian based methods interleave the independent angle and scale attributes, which affects detection performance and suppresses multi-scale detection ability. We propose a novel Scale-Aware Polar Representation (SAPR) framework to better learn arbitrary shapes and handle the scale variation of text. The decoupling learning of scale and angle attributes in Polar coordinates produces promising improvement. We then propose line IoU loss and symmetry sine loss to effectively optimize the scale and angle, respectively. The base network is equipped with multi-path decoder to better utilize the multi-scale features. A more accurate and simple center line calculation is also developed to automatically fit various scales. Extensive experiments on benchmarks and ablation study solidly demonstrate the scala-aware ability and excellent performance of SAPR.

References

1. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: European conference on computer vision, Springer (2016) 56–72
2. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2017) 5551–5560
3. Ma, J., Shao, W., Ye, H., Wang, L., Wang, H., Zheng, Y., Xue, X.: Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia **20** (2018) 3111–3122
4. Liu, J., Liu, X., Sheng, J., Liang, D., Li, X., Liu, Q.: Pyramid mask text detector. arXiv preprint arXiv:1903.11800 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
6. Huang, L., Yang, Y., Deng, Y., Yu, Y.: Densebox: Unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015)
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. (2015) 91–99
8. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision, Springer (2016) 21–37
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 3431–3440
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
11. Long, S., Ruan, J., Zhang, W., He, X., Wu, W., Yao, C.: Textsnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 20–36
12. Wang, X., Jiang, Y., Luo, Z., Liu, C.L., Choi, H., Kim, S.: Arbitrary shape scene text detection with adaptive text region representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 6449–6458
13. Tian, Z., Shu, M., Lyu, P., Li, R., Zhou, C., Shen, X., Jia, J.: Learning shape-aware embedding for scene text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4234–4243
14. Liu, Z., Lin, G., Yang, S., Liu, F., Lin, W., Goh, W.L.: Towards robust curve text detection with conditional spatial expansion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 7269–7278
15. Zhang, C., Liang, B., Huang, Z., En, M., Han, J., Ding, E., Ding, X.: Look more than once: An accurate detector for text of arbitrary shapes. arXiv preprint arXiv:1904.06535 (2019)
16. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9365–9374
17. Xu, Y., Wang, Y., Zhou, W., Wang, Y., Yang, Z., Bai, X.: Textfield: Learning a deep direction field for irregular scene text detection. IEEE Transactions on Image Processing (2019)

18. Zhu, Y., Du, J.: Textmountain: Accurate scene text detection via instance segmentation. arXiv preprint arXiv:1811.12786 (2018)
19. Zhu, Y., Du, J.: Sliding line point regression for shape robust scene text detection. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 3735–3740
20. Xue, C., Lu, S., Zhang, W.: Msr: Multi-scale shape regression for scene text detection. arXiv preprint arXiv:1901.02596 (2019)
21. Qiao, L., Tang, S., Cheng, Z., Xu, Y., Niu, Y., Pu, S., Wu, F.: Text perceptron: Towards end-to-end arbitrary-shaped text spotting. arXiv preprint arXiv:2002.06820 (2020)
22. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: Textboxes: A fast text detector with a single deep neural network. In: Thirty-First AAAI Conference on Artificial Intelligence. (2017)
23. Liao, M., Zhu, Z., Shi, B., Xia, G.s., Bai, X.: Rotation-sensitive regression for oriented scene text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5909–5918
24. He, W., Zhang, X.Y., Yin, F., Liu, C.L.: Deep direct regression for multi-oriented scene text detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2017) 745–753
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. (2017) 2961–2969
26. Lyu, P., Yao, C., Wu, W., Yan, S., Bai, X.: Multi-oriented scene text detection via corner localization and region segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7553–7563
27. Deng, D., Liu, H., Li, X., Cai, D.: Pixellink: Detecting scene text via instance segmentation. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
28. Wang, W., Xie, E., Li, X., Hou, W., Lu, T., Yu, G., Shao, S.: Shape robust text detection with progressive scale expansion network. arXiv preprint arXiv:1903.12473 (2019)
29. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2550–2558
30. Tang, J., Yang, Z., Wang, Y., Zheng, Q., Xu, Y., Bai, X.: Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. Pattern Recognition **96** (2019) 106954
31. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 12193–12202
32. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2403–2412
33. Fourure, D., Emonet, R., Fromont, E., Muselet, D., Tremeau, A., Wolf, C.: Residual conv-deconv grid network for semantic segmentation. arXiv preprint arXiv:1707.07958 (2017)
34. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia, ACM (2016) 516–520
35. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 2315–2324

36. Ch'ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Volume 1., IEEE (2017) 935–942
37. Liu, Y., Jin, L., Zhang, S., Luo, C., Zhang, S.: Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recognition **90** (2019) 337–345
38. Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z.: Detecting texts of arbitrary orientations in natural images. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2012) 1083–1090
39. Yao, C., Bai, X., Liu, W.: A unified framework for multioriented text detection and recognition. IEEE Transactions on Image Processing **23** (2014) 4737–4749
40. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255
41. Li, J., Zhang, C., Sun, Y., Han, J., Ding, E.: Detecting text in the wild with deep character embedding network. In: Asian Conference on Computer Vision, Springer (2018) 501–517
42. Liu, Y., Jin, L., Zhang, S., Luo, C., Zhang, S.: Curved scene text detection via transverse and longitudinal sequence connection. Pattern Recognition **90** (2019) 337–345
43. Chen, J., Lian, Z., Wang, Y., Tang, Y., Xiao, J.: Irregular scene text detection via attention guided border labeling. Science China Information Sciences **62** (2019) 220103
44. Dai, Y., Huang, Z., Gao, Y., Xu, Y., Chen, K., Guo, J., Qiu, W.: Fused text segmentation networks for multi-oriented scene text detection. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE (2018) 3604–3609
45. Qin, X., Zhou, Y., Yang, D., Wang, W.: Curved text detection in natural scene images with semi-and weakly-supervised learning. arXiv preprint arXiv:1908.09990 (2019)
46. Yang, Q., Cheng, M., Zhou, W., Chen, Y., Qiu, M., Lin, W., Chu, W.: Inceptext: A new inception-text module with deformable psroi pooling for multi-oriented scene text detection. arXiv preprint arXiv:1805.01167 (2018)
47. Liu, Z., Lin, G., Yang, S., Feng, J., Lin, W., Goh, W.L.: Learning markov clustering networks for scene text detection. arXiv preprint arXiv:1805.08365 (2018)
48. Ma, C., Zhong, Z., Sun, L., Huo, Q.: A relation network based approach to curved text detection. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE (2019) 707–713
49. Wang, Y., Xie, H., Zha, Z.J., Xing, M., Fu, Z., Zhang, Y.: Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 11753–11762
50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
51. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. International Journal of Document Analysis and Recognition (IJDAR) **8** (2006) 280–296
52. Xue, C., Lu, S., Zhan, F.: Accurate scene text detection through border semantics awareness and bootstrapping. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 355–372