# An Adaptive Rectification Model for Arbitrary-Shaped Scene Text Recognition

Ye Qian
mf1833053@smail.nju.edu.cn
Long Chen
mg1933003@smail.nju.edu.cn
Feng Su*
suf@nju.edu.cn

State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

**Abstract**

Recognizing scene text in natural images is challenging due to the irregular or distorted shapes of many text instances. In this paper, we propose a novel adaptive rectification model for robust recognition of arbitrary-shaped scene text. The rectification model approximates the complex non-uniform deformation required for rectifying the text with a group of localized linear projective transformations, which better preserve text's shape characteristics than non-linear deformations like TPS during the rectification. By end-to-end training with a text recognition network, the rectification model can effectively learn to transform the input text image to a more regular form that simplifies subsequent recognition. Experiment results on benchmarks demonstrate the effectiveness of the proposed rectification model for scene text recognition.

## 1 Introduction

Scene text in natural images carries rich semantic information of the image and is therefore of great value to various image applications. On the other hand, scene text often has an irregular shape such as curved or perspectively distorted, which brings significant difficulty to robustly recognizing the text from the image.

Compared to traditional character-oriented bottom-up schemes for text recognition [21, 27], recent methods [14, 25] mostly employ a word-oriented scheme which recognizes the character sequence of a word from the image using certain sequence recognition models such as the widely adopted encoder-decoder framework, and usually achieve higher recognition performance than character-oriented methods.

For recognizing arbitrary-shaped text specifically, two effective mechanisms have been proposed in recent works [15, 19, 30] — 2D attention and shape rectification. The 2D attention mechanism [15] extends 1D attention to better depict the spatial layout of one text and helps extract accurate features of characters for recognition. Comparatively, the text shape rectification mechanism [19, 26] transforms the input text to a more regular shape, so


* Corresponding author

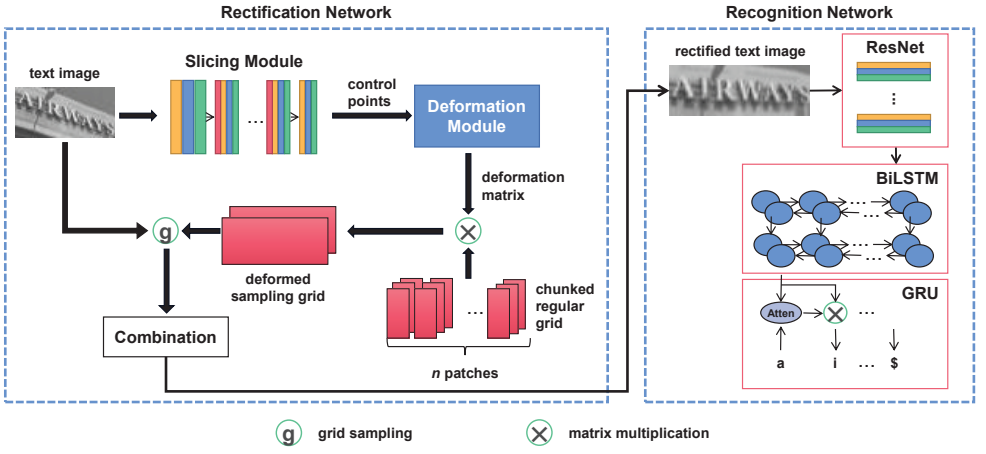


Figure 1: The architecture of the proposed rectification-based scene text recognition model.

that the rectified text can be recognized with common techniques designed to handle regular text.

In this paper, we propose a novel rectification model to adaptively regularize the shape of a scene text to facilitate subsequent recognition. Most of previous rectification methods [24, 26] employed thin plate spline (TPS) transform with spatial transform network (STN) [10] for text shape deformation, which provides great deformation flexibility but may also distort character shape in an unwanted way (e.g., bending a straight stroke) due to its non-linearity. Comparatively, our work employs group of linear transformations for text rectification to reduce distortions to the text's shape while still providing sufficient deformation flexibility. Moreover, compared to the method [19] which rectified a text image by predicting respective offsets of fixedly divided image patches, our work employs an adaptive non-uniform grid for slicing patches and a projective deformation for each patch, which are more flexible and accurate than the rectification scheme in [19].

Figure 1 shows the architecture of our rectification-based text recognition model, which comprises two main building blocks: an adaptive text shape rectification module employing a slicing, deforming, and recombining pipeline and an attention-based recognition module. The main contributions of our work can be summarized as follows:

- We propose a novel text shape rectification model for robust recognition of arbitrary-shaped scene text. Different from previous rectification methods based on non-linear transformations, our model employs a group of localized linear projective transformations to approximate the complex non-uniform transformation required for rectifying irregular text while better preserving text's shape characteristics.

- We further introduce global smoothness constraint on neighboring local transformations by sampling control points of transformation from predicted Bezier boundaries of text, which keeps the rectification model from introducing unnatural deformations.

- Through end-to-end training with a text recognition network, the proposed rectification model effectively improves the performance of the whole text recognition model on benchmarks in the experiments.

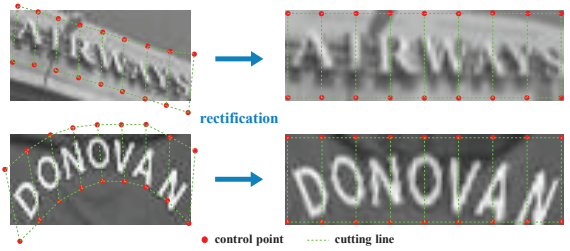


Figure 2: Illustration of adaptively slicing a text region into non-uniform patches for rectification based on predicted control points and cutting lines.

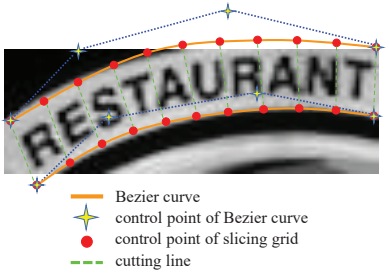


Figure 3: Illustration of Bezier curves for sampling control points of the slicing grid.

## 2 Text Rectification through Slicing, Deformation and Recombination

We propose a novel and effective rectification model for arbitrary-shaped text, which requires no extra label information and therefore can be inserted prior to a recognition network and learn to adaptively transform irregular text to a favorable form for subsequent recognition through end-to-end training in a weak supervision way with common word annotations only. Specifically, as one key requirement for a text shape rectification model is to keep characters from arbitrary and destructive changes in their shapes, our rectification model employs a group of localized linear projective transformations (instead of non-linear transforms like TPS) to better preserve a character's shape characteristics for accurate recognition.

As shown in Fig. 1, our text rectification model is composed of a slicing module, a deformation module, and a recombination module. The network first adaptively slices a text region into a series of quadrilateral patches, then deforms them into regular rectangular target patches, and finally combines rectified patches into a new text region with regular linear layout.

### 2.1 Slicing Module

Given a text image, the slicing module first locates and adaptively slices the arbitrary-shaped text region into a strip of consecutive small patches which will be deformed individually during rectification. We propose two slicing schemes to obtain the appropriate position of each patch.

In the first basic scheme, as shown in Fig. 2, we predict the coordinates of $N+1$ pairs of control points $(\mathbf{p}_i^0, \mathbf{p}_i^1)_{i=0..N}$ of a slicing grid covering the text region, using a convolutional subnetwork (denoted by LN) which cascades six convolution blocks (interleaved with five max pooling layers) and two fully connected layers. Each pair of control points defines a cutting line, and every two neighbouring cutting lines define a quadrilateral patch, so that the whole text region is adaptively sliced into a series of $N$ adjacent patches $\{S_i\}_{i=0..N}$.

Generally, the smaller are the sliced patches, the more smoothly a curved text contour is approximated by the set of quadrilateral patches. However, as our rectification model is weakly supervised with only word labels (i.e., without positional annotations), it is usually hard for LN to directly and accurately predict a set of dense control points of the slicing grid.

Accordingly, as the second slicing scheme, we propose a two-step slicing mechanism which generates a dense slicing grid and ensures better smoothness of the patches.

### 2.1.1 Dense Slicing of Text Region via Bezier Sampling

To obtain a dense slicing grid for a text region, we first approximate the upper and lower edges of the text region with two *cubic Bezier curves*, as shown in Fig. 3. To avoid directly predicting the control points $\mathbf{CP} = \{\mathbf{cp}_i\}_{i=0..3}$ of the Bezier curves, which is relatively difficult with weak supervision, we use the LN subnetwork to predict the coordinates of a set of $M+1$ sampling points $\mathbf{RP} = \{\mathbf{rp}_0, \mathbf{rp}_1, \ldots, \mathbf{rp}_M\}$ on each Bezier curve, which are supposed to satisfy:

$$\mathbf{RP} = \mathbf{V} \cdot \mathbf{CP} \tag{1}$$

$$\mathbf{V} = \begin{bmatrix} (1-t_0)^3 & 3(1-t_0)^2 t_0 & 3(1-t_0)t_0^2 & t_0^3 \\ (1-t_1)^3 & 3(1-t_1)^2 t_1 & 3(1-t_1)t_1^2 & t_1^3 \\ (1-t_2)^3 & 3(1-t_2)^2 t_2 & 3(1-t_2)t_2^2 & t_2^3 \\ \cdots & \cdots & \cdots & \cdots \\ (1-t_M)^3 & 3(1-t_M)^2 t_M & 3(1-t_M)t_M^2 & t_M^3 \end{bmatrix} \tag{2}$$

where $t_i$ is the Bezier curve variable corresponding to the $i$th sampling point $\mathbf{rp}_i$, which is computed as $t_i = sd_i/sd_M$, while $sd_i$ denotes the approximate length of the curve segment from $\mathbf{rp}_0$ to $\mathbf{rp}_i$ and is calculated by accumulating distances between each pair of adjacent sampling points.

We then employ the standard least square method to estimate the most suitable solution $\widehat{\mathbf{CP}}$ for the Bezier curve's control points based on the predicted sampling points $\mathbf{RP}$, under the constraint of minimizing $\|\mathbf{X}\,\mathbf{CP} - \mathbf{RP}\|$:

$$\widehat{\mathbf{CP}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{RP} \tag{3}$$

where $\mathbf{X}$ denotes the Bernstein Polynomials [17] matrix of the Bezier curve.

Finally, given the upper and lower Bezier curve edges of the text region, we sample two series of dense control points of the slicing grid on the two edges respectively, using a set of $N+1$ uniformly spaced Bezier curve variable values $t$ in the range $[0,1]$, which enforce the similar sizes of different patches as well as the smoothness of the rectified text region.

## 2.2 Deformation Module

We model the mapping from a quadrilateral source patch to the corresponding rectangular target patch in the rectified text region with a linear projective transformation, which better preserves shape characteristics of characters than non-linear transformations like TPS.

A 2D projective transformation is a linear transformation between two 2D points $\mathbf{x}$ and $\mathbf{x}'$ (in homogeneous coordinates) and can be represented by a non-singular $3 \times 3$ matrix $\mathbf{H}$:

$$\mathbf{x}' = \mathbf{H}\mathbf{x}' \qquad \mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{4}$$

As $\mathbf{H}$ is a homogeneous matrix having eight degrees of freedom, given the four pairs of corresponding control points (i.e., the vertices of the source and target patches), we can set $h_{33} = 1$ and use the Direct Linear Transformation (DLT) algorithm to estimate $\mathbf{H}$ [7].

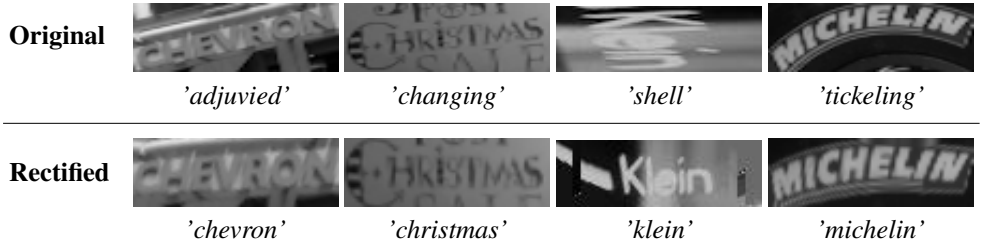


Figure 4: Illustration of text shape rectification. Row 'Original' presents input text images. Row 'Rectified' presents corresponding rectified images. The text under an image is the corresponding recognition result.

Based on the obtained deformation matrix $\mathbf{H}$, we generate the rectified target patch using grid sampling. Specifically, for one pixel with homogeneous coordinates $\mathbf{p}_t$ in the target patch, it is mapped back to a position $\mathbf{p}_s$ in the source patch:

$$\mathbf{p}_s = \mathbf{H}^{-1}\mathbf{p}_t \tag{5}$$

The target pixel's value can then be computed based on the pixels neighbouring to the position $\mathbf{p}_s$ in the source patch using bilinear interpolation.

### 2.3 Recombination Module

Given the deformed patches, the recombination module joins the patches together along the width direction to form a rectified text image to be fed to the recognition module. In this work, the rectification network takes an input text image of size $32 \times 64$ and produces an output rectified image of size $32 \times 100$.

Figure 4 illustrates several examples of rectified text images, along with the original input text images and the recognition results on both original and rectified images. It can be seen that, through end-to-end training with the recognition network, our rectification model effectively learns to deform the text image in a way leading to improved recognition results.

## 3 Text Recognition

On the basis of the rectification model regularizing the text shape to a form easier for recognition, we employ a light-weight, attention-based encoder-decoder network for text recognition, which is essentially the same as those used in [19, 26]. As shown in Fig. 1, the recognition network consists of three parts: the ResNet backbone, the BiLSTM sequential encoder, and the attention-based gated recurrent unit (GRU) decoder.

Specifically, a 45-layer residual network is first employed to extract features from a rectified text image, which generates a feature map composed of 512 channels with a height of 1. Next, the feature map is divided into a sequence of column feature vectors, and two cascaded BiLSTMs (each with a hidden layer consisting of 256 units) are employed to capture sequential long-range dependencies between input feature vectors with the hidden states of both forward and backward directions, which are concatenated to form an output sequence of feature vectors $\{\mathbf{h}_i\}_{i=1..L}$. An attentive GRU decoder is then employed to predict a se-

quence of character label distribution vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, where $N \leqslant T$ and $T$ denotes the largest number of steps processed by the decoder.

### 3.1 Loss Function

As the rectification model does not require additional supervision information, the loss of the whole text recognition model is formulated as:

$$Loss = \sum_{i=1}^{N} \sum_{j=1}^{M} \mathbb{I}(\hat{\mathbf{y}}_i^j = 1) log(\mathbf{y}_i^j) \tag{6}$$

where $N$ is the length of the predicted character label distribution sequence $\{\mathbf{y}_i\}$, $M$ is the total number of different characters, $\{\hat{\mathbf{y}}_i\}$ is the one-hot ground-truth distribution sequence, and $\mathbb{I}(\cdot)$ is a binary function that returns 1 if its input is evaluated to true and 0 otherwise.

## 4 Experiments

### 4.1 Datasets

We adopt seven standard scene text recognition benchmarks in the experiments, including three irregular text datasets **SVT-P** [22], **CUTE80** (CT80) [23], and **ICDAR 2015** (IC15) [13], and four regular text datasets **IIIT5K-Words** (IIIT5K) [20], **Street View Text** (SVT) [27], **ICDAR 2003** (IC03) [18], and **ICDAR 2013** (IC13) [12]. Details of these datasets can be found in [26]. We employ *word recognition accuracy* as the recognition performance metric, which is defined as $\frac{|C|}{|T|}$ with $C$ and $T$ being the set of correctly recognized words and ground-truth words, respectively.

### 4.2 Implementation Details

We implement the proposed scene text recognition network with PyTorch framework and conduct the experiments on a NVIDIA Tesla V100 GPU.

We train our text recognition network end to end on the 8-million synthetic data released by Jaderberg *et al.* [8] and the 6-million synthetic data released by Gupta *el al.* [6], using only word annotations. The Adadelta optimizer is employed in the training with minibatches of size 64. The learning rate is initially set to 1.0 for the first three epochs and is adjusted to 0.1 for the two subsequent epochs.

For a test image whose height is larger than the width, we rotate it by 90 degrees both clockwise and anticlockwise for testing. Prior to being fed into the recognition network, the width and height of an image are resized to 200 and 64. The recognition network achieves about 38.5 FPS in the inference. No lexicon information is exploited in all experiments.

### 4.3 Ablation Study

#### 4.3.1 Effectiveness of Text Rectification Model

We verify the effect of the proposed rectification model on enhancing the final text recognition accuracy by comparing the recognition performance with some variant rectification

Table 1: Effectiveness of the proposed rectification model for text recognition

| | IC15 | SVT-P | CT80 | SVT |
|---|---|---|---|---|
| Baseline | 71.1 | 74.8 | 71.8 | 86.0 |
| MORN | 73.9 | 79.7 | 81.9 | 88.1 |
| TPS | 76.1 | 78.5 | 79.5 | 89.5 |
| **Proposed** | **77.1** | **81.3** | **82.9** | **89.7** |

Table 2: Comparison of recognition performance using varied numbers $N$ of patches in the rectification

| $N$ | IC15 | SVT-P | CT80 |
|---|---|---|---|
| 1 | 75.1 | 81.0 | 80.2 |
| 10 | 76.3 | 80.3 | 80.9 |
| 20 | 76.5 | 80.9 | 81.9 |
| **25** | **77.1** | **81.3** | **82.9** |
| 50 | 74.4 | 80.3 | 79.5 |

models in Table 1. Specifically, the model "Baseline" removes the rectification module, feeding the input text image directly to the recognition network. The model "MORN" replaces the proposed rectification module with the MORN rectification network proposed in [19]. The model "TPS" employs the rectification method proposed in [26], which uses the TPS transformation with STN for text rectification and an essentially same recognition network as ours.

Compared to the baseline model, introducing the proposed rectification module significantly enhances the performance of the whole recognition model on all datasets, including the other three regular datasets IC03 (+0.7%), IC13 (+1.8%), and IIIT5K (+2.0%) not shown in Table 1, showing the proposed rectification model successfully learns to spatially transform the input text image to a favorable form for recognition during end-to-end training.

Compared to the MORN rectification network [19], the proposed rectification model exhibits better rectification capability by predicting an adaptive, arbitrary-shaped sampling grid, which more accurately and flexibly depicts the potential deformations of text patches compared to the fixed grid used in MORN. Compared to the TPS-based rectification method in [26], the proposed rectification model more effectively improves the text image for subsequent recognition by restricting the deformations of text region to linear ones to avoid introducing harmful non-linear distortions to character shapes while still providing sufficient deformation flexibility.

Figure 5 presents several examples of rectified text images by MORN [19], TPS-based [26], and the proposed rectification models respectively, along with corresponding recognition results. The more accurate recognition results obtained with the proposed rectification model on rectified text images show its advantages for text shape rectification.

#### 4.3.2 Influence of Different Numbers of Patches in Rectification

We investigate the impact of slicing a text region into different numbers of patches for rectification on the recognition performance in Table 2. As the rectified image in this work has a fixed width of 100 pixels, we limit the number of patches to those shown in the table, so that each sliced patch has an integer width.

As shown by the results, slicing a text region into too few patches (e.g., $\leq 10$) will produce large patches, which are generally difficult to accurately depict the non-uniform deformation required to rectify distortions of a text. As the number of patches increases, the contour of a curved text can be more accurately approximated by the denser control points of local projective transformations, resulting in improved rectification and recognition results. However, as the size of the rectified image is fixed, slicing the text region into too many patches will result in too small patches, which will eventually cause the drop of performance.

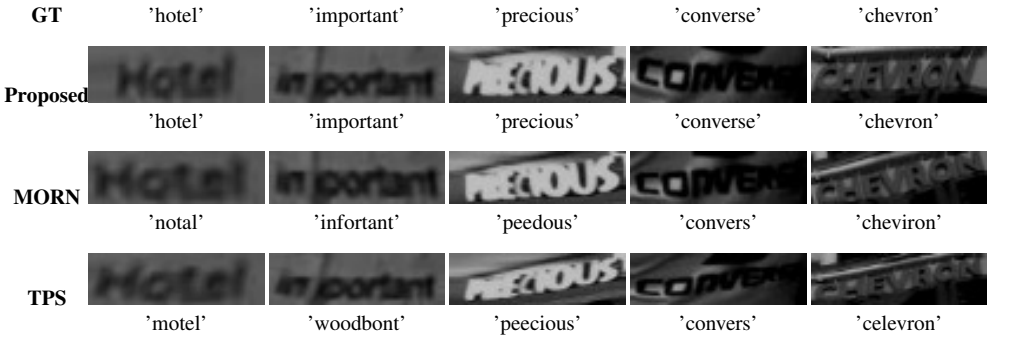


Figure 5: Examples of rectified text images by the proposed, MORN, and TPS-based rectification models and corresponding recognition results (text below images). GT denotes the ground truth.


Figure 6: Illustration of the effect of Bezier curve-based dense sampling of slicing grid points on the rectification result. The crosses in the original image indicate the control points of the slicing grid.

### 4.3.3 Effectiveness of Bezier Curve-Based Sampling of Slicing Grid

Table 3 compares the recognition performance employing the basic (directly predicting slicing grid points) and the Bezier curve sampling-based slicing schemes in the rectification respectively, and Figure 6 shows two examples of rectification results by the two variant schemes.

Table 3: Comparison of recognition performance with the basic and the Bezier curve sampling-based slicing schemes for rectification

| | IC15 | SVT-P | CT80 |
|---|---|---|---|
| Basic | 76.9 | 80.3 | 81.9 |
| Bezier | **77.1** | **81.3** | **82.9** |

It can be seen that the Bezier curve-based sampling mechanism, which helps generate a smoother slicing grid with denser grid points, improves the rectification results (e.g. more consistent deformations of different text parts) and enhances the final recognition accuracy.

## 4.4 Comparison with State-of-the-Art Methods

We compare our method with some state-of-the-art scene text recognition methods in Table 4. Note that, as the main goal of this work is an effective text shape rectification model for robust recognition of arbitrary-shaped text, we employed a common, light-weight text recognition network and did not exploit any additional datasets nor data augmentation schemes for training the model.

**Irregular Text Recognition.** Our method achieves the top recognition accuracy on SVT-P and IC15, which comprise large numbers of perspectively distorted, arbitrarily-oriented, and curved text, among methods exploiting only word-level annotations of text. Our method also obtains the third best recognition accuracy on CT80 dataset composed of many curved

Table 4: Recognition accuracy on regular and irregular datasets without lexicon. In each column, the best performing result is shown in bold font, and the second best result is shown with underline. Note the approaches marked with * were trained with additional character-level annotations besides standard word-level annotations and therefore not included in ranking.

| Method | Irregular Text | | | Regular Text | | | |
|---|---|---|---|---|---|---|---|
| | IC15 | SVT-P | CT80 | IIIT5K | SVT | IC13 | IC03 |
| Bissacco *et al.* [2] | - | - | - | - | 78.0 | 87.6 | - |
| Jaderberg *et al.* [9] | - | - | - | - | 71.7 | 81.8 | 89.6 |
| Jaderberg *et al.* [11] | - | - | - | - | 80.7 | 90.8 | 93.1 |
| Shi *et al.* [25] | - | 66.8 | 54.9 | 78.2 | 80.8 | 86.7 | 93.1 |
| Shi *et al.* [24] | - | 71.8 | 59.2 | 81.9 | 81.9 | 88.6 | 90.1 |
| Lee *et al.* [14] | - | - | - | 78.4 | 80.7 | 90.0 | 88.7 |
| Liu *et al.** [16] | 74.2 | 78.9 | - | 92.0 | 85.5 | 91.1 | 92.0 |
| Yang *et al.** [30] | - | 75.8 | 69.3 | - | - | - | - |
| Cheng *et al.** [3] | 70.6 | - | - | 87.4 | 85.9 | 93.3 | 94.2 |
| Bai *et al.** [1] | 73.9 | - | - | 88.3 | 87.5 | 94.4 | 94.6 |
| Cheng *et al.* [4] | 68.2 | 73.0 | 76.8 | 87.0 | 82.8 | - | 91.5 |
| Luo *et al.* [19] | 68.8 | 76.1 | 77.4 | 91.2 | 88.3 | 92.4 | **95.0** |
| Shi *et al.* [26] | 76.1 | 78.5 | 79.5 | 93.4 | 89.5 | 91.8 | 94.5 |
| Li *et al.* [15] | 69.2 | 76.4 | 83.3 | 91.5 | 84.5 | 91.0 | - |
| Zhan *et al.* [32] | 76.9 | 79.6 | 83.3 | 93.3 | **90.2** | 91.3 | - |
| Yang *et al.** [29] | 78.7 | 80.8 | 87.5 | 94.4 | 88.9 | 93.9 | 95.0 |
| Wang *et al.* [28] | 74.5 | 80.0 | **84.4** | **94.3** | 89.2 | **93.9** | **95.0** |
| **Ours** | **77.1** | **81.3** | 82.9 | 93.4 | 89.7 | 93.4 | 94.3 |

text. Particularly, compared to previous rectification-based methods [19, 24, 26, 32], our method achieves averagely higher accuracy with a common recognition network similar to those used in these methods. Note the method [29], which employs TPS transformation for rectifying text regions, relies on additional, stronger supervision of annotations of character-level geometrical attributes and the segmentation mask of the text center line. The results demonstrate the effectiveness of the proposed text shape rectification model for irregular text recognition.

Different from our work that focuses on the shape rectification mechanism for irregular text recognition, many of latest scene text recognition methods focused on enhancing the recognition network with sophisticated semantic and language models, which helped to rectify incorrect recognition results (e.g., caused by irregular text shape, blurring, or noises) exploiting the semantic context information and significantly enhanced the overall text recognition accuracy. For instance, SRN [31] introduced a global semantic reasoning module (GSRM) into the text recognition model, which captured the global semantic context through multi-way parallel transmission. Combining GSRM with a parallel visual attention module and a visual-semantic fusion decoder, SRN achieved 82.7, 85.1, and 87.8 accuracy on irregular text benchmarks IC15, SVT-P, and CT80 respectively. Furthermore, ABINet [5] proposed an autonomous, bidirectional cloze network with bidirectional feature representation and an iterative correction mechanism for the language model, and achieved 86.0 on IC15, 89.3 on SVT-P, and 89.2 on CT80 (ABINet-LV). As an end-to-end trainable module, our text shape rectification model can be combined with these sophisticated recognition

models for improved recognition performance as shown in Section 4.5.

**Regular Text Recognition.** Our rectification-based text recognition method also achieves competitive performance on regular text datasets, showing the adaptive rectification mechanism avoids adverse effects on text with relatively regular shapes and is useful for improving the recognition accuracy.

## 4.5 Combine Rectification Network with Other Recognition Model

To inspect the synergy of our proposed text rectification mechanism with more recent recognition models than the light-weight one presented in Section 3, we combine our rectification network with two variant recognition models of ABINet [5]. The first variant 'Baseline-1' employs the vision model of ABINet, which consists of a ResNet+Transformer backbone network and a position attention module for character probability prediction. The second variant 'Baseline-2' further integrates the language model of ABINet based on a bidirectional cloze network (BCN) with Baseline-1. We insert the rectification network as a front module before the vision model of two baselines to create two combined recognition models, and train all models on the datasets as described in Sections 4.1 and 4.2.

Tables 5 and 6 compare the recognition performance of the baselines and the combined models. It can be seen that, compared to the baselines, introducing the proposed rectification module contributes an improvement in recognition accuracy on the benchmarks. For example, Combined-1 model outperforms Baseline-1 by 2.9% on IC15, 1.1% on SVT-P, 1.4% on IC13, and 1.6% on SVT, and Combined-2 model outperforms Baseline-2 by 1.5% and 1.2% on IC15 and SVT respectively. Note the performance improvement by introducing the rectification module to Baseline-2 (which employs a sophisticated language model) is smaller than to Baseline-1. One probable reason is that the language model pre-trained on large text corpus in Baseline-2 effectively corrected many of character recognition errors caused by irregular text shapes.

Table 5: Recognition performance of the combination of the proposed rectification network with the vision model of ABINet ('Baseline-1')

| | IC15 | SVT-P | IC13 | SVT |
|---|---|---|---|---|
| Baseline-1 | 78.7 | 81.5 | 92.9 | 87.3 |
| **Combined-1** | **81.6** | **82.6** | **94.3** | **88.9** |

Table 6: Recognition performance of the combination of the proposed rectification network with the vision & language models of ABINet ('Baseline-2')

| | IC15 | SVT-P | IC13 | SVT |
|---|---|---|---|---|
| Baseline-2 | 82.8 | 88.8 | 96.9 | 91.2 |
| **Combined-2** | **84.3** | **89.0** | **97.1** | **92.4** |

# 5 Conclusions

We present a novel text shape rectification model for recognizing arbitrary-shaped scene text. The rectification model employs an effective slicing, deforming, and recombining pipeline to transform the shape of an input text to a more favorable form for subsequent recognition, which enhances the performance of the whole recognition model through end-to-end training. Experiment results demonstrate the effectiveness of our rectification-based text recognition method.

# References


[1] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1508–1516. IEEE Computer Society, 2018.

[2] Alessandro Bissacco, Mark Cummins, Yuval Netzer, and Hartmut Neven. PhotoOCR: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision ICCV 2013*, pages 785–792, Dec 2013.

[3] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5086–5094, Oct 2017.

[4] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. AON: towards arbitrarily-oriented text recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5571–5579, 2018.

[5] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7098–7107, 2021.

[6] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR 2016*, pages 2315–2324, June 2016. doi: 10.1109/CVPR.2016.254.

[7] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2003.

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *CoRR*, abs/1406.2227, 2014.

[9] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pages 512–528, 2014.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[11] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[12] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, avid Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. ICDAR 2013 robust reading competition. In *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*, pages 1484–1493, 2013.

[13] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, iri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 1156–1160, 2015.

[14] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for OCR in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2231–2239, June 2016.

[15] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8610–8617, Jul 2019.

[16] Wei Liu, Chaofeng Chen, Kwan-Yee Kenneth Wong, Zhizhong Su, and Junyu Han. STAR-Net: A spatial attention residue network for scene text recognition. In *BMVC*, 2016.

[17] George G. Lorentz. *Bernstein Polynomials*. AMS Chelsea Publishing, second edition, 1986.

[18] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, and Robert Young. ICDAR 2003 robust reading competitions. In *7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK*, pages 682–687, 2003.

[19] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019.

[20] Anand Mishra, Karteek Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, pages 1–11, 2012.

[21] Lukas Neumann and Jiri Matas. Real-time scene text localization and recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3538–3545, 2012.

[22] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 569–576, 2013.

[23] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014.

[24] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4168–4176, Jun 2016.

[25] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, Nov 2017.

[26] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2035–2048, Sep. 2019.

[27] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV 2011*, pages 1457–1464, Nov 2011. doi: 10.1109/ICCV.2011.6126402.

[28] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12216–12224, Apr 2020.

[29] Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao, and Xiang Bai. Symmetry-constrained rectification network for scene text recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9146–9155, 2019.

[30] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C. Lee Giles. Learning to read irregular text with attention mechanisms. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3280–3286, 2017.

[31] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12110–12119, 2020.

[32] Fangneng Zhan and Shijian Lu. ESIR: End-to-end scene text recognition via iterative image rectification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2054–2063, Jun 2019.