



# Adversarial learning based attentional scene text recognizer<sup>☆</sup>

Jinyuan Zhao<sup>a,b</sup>, Yanna Wang<sup>a</sup>, Baihua Xiao<sup>a,\*</sup>, Cunzhao Shi<sup>a</sup>, Jingzhong Jiang<sup>a,b</sup>, Chunheng Wang<sup>a</sup>

<sup>a</sup> Institute of Automation, Chinese Academy of Sciences (CASIA), 95 Zhongguancun East Road, Beijing, 100190, PR China

<sup>b</sup> University of Chinese Academy of Sciences (UCAS), No. 19 (A) Yuquan Road, Shijingshan District, Beijing, 100049, PR China

## ARTICLE INFO

### Article history:

Received 16 February 2020

Revised 20 May 2020

Accepted 17 July 2020

Available online 18 July 2020

### Keywords:

Scene text recognition

Generative adversarial network

Image rectification

## ABSTRACT

In this paper, we propose an adversarial learning based attentional scene text recognizer to solve the distortion problem of scene text image. We choose a rectification module which can rectify images in both horizontal and vertical directions, and use a recognizer based on the attention mechanism. Through the adversarial learning of the rectification network and the recognition network, we iteratively improve the rectification effect and the recognition performance. The entire network is trained with weak supervision, so only images and corresponding text labels are needed. Our method achieves high performance for both regular and irregular scene text images, and the experimental results tested on multiple benchmarks prove that our method achieves the performance of state-of-the-art.

© 2020 Elsevier B.V. All rights reserved.

## 1. Introduction

Texts can bring us great help to understand the content of an image. The research on text recognition can promote the development of many intelligent products, such as scene image retrieval, automatic driving, product recognition and so on. With the development of deep learning technology and scene text detection method, the method of scene text recognition has received more attention.

It is worth noting that scene text recognition is a challenging research topic due to the various shapes and distorted patterns of irregular texts, as shown in Fig. 1. In this paper, we propose the Adversarial Learning based Attentional Scene Text Recognizer (ASTR), which can read rotated, scaled and stretched characters in different scene texts.

The ASTR consists of a rectification network and a recognition network. Given an irregular scene text image, we first obtain the regular image through our rectification network and then use the recognizer for text recognition.

Inspired by the outstanding performance of the generative adversarial network [10], we introduce the strategy of adversarial training into the scene text recognition task. We regard the rectification network and the recognition network as the generator

and the discriminator of adversarial learning, respectively. However, different from the traditional GAN framework, the discriminator in ASTR does not directly judge the rectified image, and the final result is obtained through the combination of our generator and discriminator.

We use the difference of recognition results of the image before and after rectification to evaluate the two sub-networks. For the rectification network, we hope that the rectified image can have better recognition result than the original image, so the rectification network seeks to maximize the difference of the recognition results. For the recognition network, we hope it can accurately identify texts in both original and rectified images, and the difference should be smaller. Through adversarial training, the rectification network and recognition network can be improved synchronously.

We have carried out extensive experiments on a series of standard datasets to prove that ASTR can obtain superior recognition performance on both regular and irregular text images. The main contributions of this paper are as follows: (1) We apply adversarial training strategy on scene text recognition and propose a proper architecture for this task; (2) Our training is weakly supervised and the effect of both rectification network and recognition network has been improved after training; (3) Comprehensive experimental results on multiple datasets demonstrate that our method achieves promising performance in terms of regular and irregular scene text image recognition compared with many traditional and state-of-the-art algorithms.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3 details

<sup>☆</sup> Handle by Associate Editor Imran Siddiqi.

\* Corresponding author.

E-mail addresses: [zhaojinyuan2016@ia.ac.cn](mailto:zhaojinyuan2016@ia.ac.cn) (J. Zhao), [wangyanna2013@ia.ac.cn](mailto:wangyanna2013@ia.ac.cn) (Y. Wang), [baihua.xiao@ia.ac.cn](mailto:baihua.xiao@ia.ac.cn) (B. Xiao), [cunzhao.shi@ia.ac.cn](mailto:cunzhao.shi@ia.ac.cn) (C. Shi), [jiangzhong2018@ia.ac.cn](mailto:jiangzhong2018@ia.ac.cn) (J. Jiang), [chunheng.wang@ia.ac.cn](mailto:chunheng.wang@ia.ac.cn) (C. Wang).



Fig. 1. Some examples of scene text images.

the proposed method. Experimental results and analysis are given in Section 4, and conclusions are presented in Section 5.

## 2. Related work

Many researchers focus on scene text recognition and put forward effective methods [6,23,25,33]. However, scene text with irregular shapes, such as perspective and curved text, is still very challenging, so more and more researchers are engaged in the field of scene text recognition. Ye and Doermann [44], Zhu et al. [48] and Baek et al. [2] these articles provide an overview of the main progress in the field of scene text detection and recognition.

The traditional text recognition method regards text recognition as a special classification task and studies the text recognition problem based on the general target recognition methods [36,40]. For instance, Bissacco [4] applied a network with five hidden layers for character classification. Using convolutional neural networks (CNNs), Jaderberg et al. [15] proposed their method for unconstrained recognition.

With the continuous optimization of the recurrent neural networks, the encoding network can better combine the context information, and methods based on sequence recognition have been further improved. Graves et al. [12] proposed an end-to-end text recognition framework, which combines CNNs and RNNs to extract and encode text features, and uses CTC loss to compare prediction results and text labels, significantly improving the trainability and recognition accuracy of text line recognition. Furthermore, the attention mechanism [3] is able to focus on information-rich areas for better recognition performance. Su et al. [37] and Lee et al. [23] combined RNNs and attention machines, and achieved great success while dealing with horizontal or slightly distorted texts in scenes.

In recent years, many researchers aim at recognizing irregular scene text with bending and distortion. Cheng et al. [7], Li et al. [24] encode the image features in both horizontal and vertical directions, so as to extract the two-dimensional features of the irregular text and achieve better recognition accuracy. Some other methods such as Shi et al. [35], Zhan and Lu [46] and Luo et al. [27] try to add a rectification network before the recognition network, firstly rectifying the irregular image to a regular text image, and then recognizing the texts in the rectified image.

All of the above strategies have significant contributions to the task of scene text recognition. Because of the complexity of the real scene and the diversity of the text distribution, it is an open problem to recognize the text in the scene. In order to make the proposed method more robust, we try to use the adversarial training strategy which has been playing an important role in image generation task, especially for the input which is not involved in training samples.

We propose an ASTR recognition framework, which uses the combination of rectification and recognition network to recognize scene text images. The rectification network acts as a generator in our adversarial learning, and the recognition network is equivalent to the discriminator. By designing the appropriate loss function and training strategy, the loss can be directly fed back to the

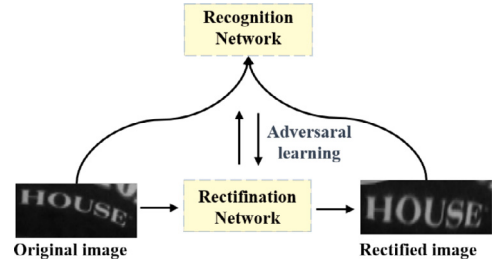


Fig. 2. The training strategy of ASTR.

rectification network and recognition network, and the rectification effect and recognition performance can be improved at the same time.

## 3. Methodology

The proposed ASTR consists of two components: one is the rectification network, the other is the recognition network. The whole network is trained by weak supervision and does not need character level annotation. In this section, we will explain our adversarial training strategy and loss design in detail, and introduce the structure of these two networks respectively.

### 3.1. Adversarial training strategy

The flowchart of ASRA is shown in Fig. 2. In this section, we will specifically introduce how to apply the adversarial learning strategy to optimize our rectification network and recognition network. In our framework, we use the rectification network as a generator, and the recognition network acts as a discriminator. By comparing the recognition results of the original image and the rectified image, the two modules are optimized iteratively.

During implementation, the rectification network and recognition network are trained through the following minimax game:

$$\arg \min_D \max_G L_{GAN}(G, D) \quad (1)$$

Here,  $G$  denotes the rectification network which acts as a generator.  $D$  denotes the recognition network which acts as a discriminator. The objective function  $L_{GAN}(G, D)$  is the cross-entropy loss of recognition results between the original image and rectified image:

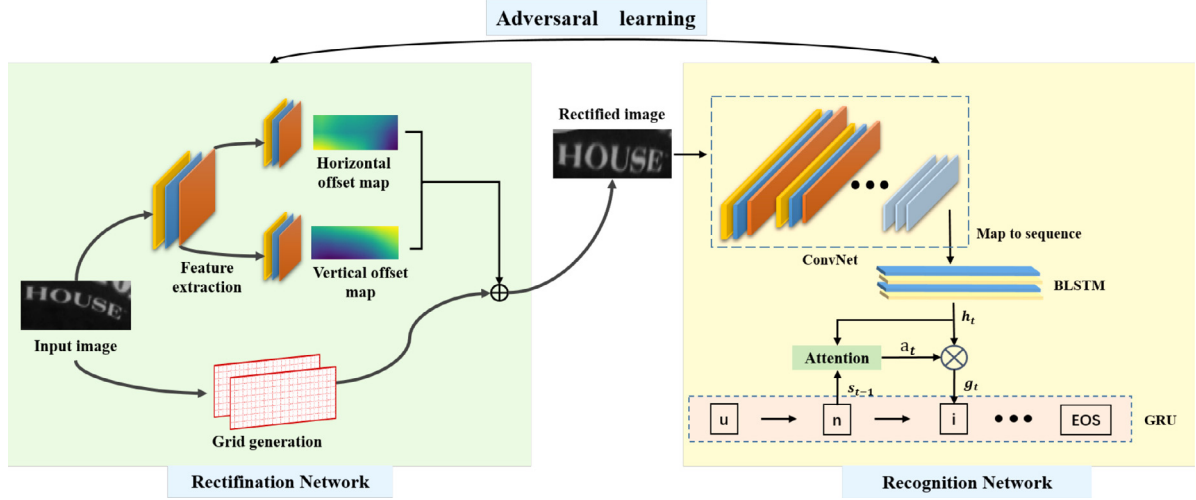
$$L_{GAN}(G, D) = E_{I \in P_{data}} \left[ - \sum_{t=1}^T (y_{ti} | I) \log(y_{ti'} | I') \right] \quad (2)$$

where  $I$  is an original scene text image from datasets  $P_{data}$ .  $I' = G(I)$ , denotes the rectified image after rectification network.  $y_{ti}$  and  $y_{ti'}$  denote the  $t$ th predicted character in the input image  $I$  and  $I'$ .

In adversarial learning, we hope to make the irregular text images accurately recognized by the recognizer after being rectified. So we calculate the cross-entropy loss between the recognition results of the rectified image and the original image. The larger the loss, the greater the difference between the rectified image and the original image, and the more obvious the rectifying effect of the rectification network on the original image.

For the recognition network, the goal is to have a higher recognition rate for all input images, that is, both the original and the rectified image should be accurately recognized. Therefore, the recognition network expects the loss calculated by Eq. (2) to be as small as possible.

At the same time, for some regular original images, we hope that the rectified image will not make them degenerated, so we add the recognition loss of rectified image ( $Loss_R$ ) to the final loss



**Fig. 3.** The network structure of ASTR. After the adversarial training, the rectification and recognition networks are combined to get the final results. We use the STN [18] based grid sampling to rectify the irregular image, and recognize the rectified image sequence-to-sequence using an attention mechanism based decoder.

of ASTR:

$$Loss_R = - \sum_{t=1}^T \log P(Y_t | I') \quad (3)$$

where  $Y_1, Y_2, \dots, Y_T$  denote the groundtruth text represented by a character sequence.

Therefore, for the rectification network, its complete loss function is as follows:

$$Loss_G = E_{I \in P_{data}} \left[ \sum_{t=1}^T (y_{ti} | I) \log(y_{ti'} | I') \right] + \lambda Loss_R \quad (4)$$

The final loss function of the recognition network is as follows:

$$Loss_D = E_{I \in P_{data}} \left[ - \sum_{t=1}^T (y_{ti} | I) \log(y_{ti'} | I') \right] + \lambda Loss_R \quad (5)$$

Here we set  $\lambda = 1$  for all our experiments. In the testing process, recognition network only recognizes rectified images. The specific network structure is shown in Fig. 3.

### 3.2. Rectification network

In the adversarial learning framework, the rectification network is equivalent to a generator, whose input is an irregular scene text image, and the output is a rectified regular text image. Inspired by Zhao et al. [47] and some excellent irregular text image recognition work [17,27,34], we choose to calculate the offset of each pixel in the irregular image from the corresponding position in the regular image to rectify the input image.

Our rectification network is based on STN [18]. The main idea is to treat the spatial transformation of the image as a learnable module. In order to compare the experimental results more fairly, we normalize the input image  $I$  to  $32 \times 100$ , consistent with most of the existing text recognition methods. We use FCN network to learn the pixel offset. The specific network structure is shown in Table 1.

Our rectification network predicts the position offsets rather than the categories of characters. The character details for classification are not necessary. We hence place a pooling layer before the convolutional layer to avoid noise and reduce the amount of calculation. Meanwhile, the output parameter  $\theta$  is a  $2 \times 8 \times 25$  feature map, wherein the first  $8 \times 25$  feature layer represents the

**Table 1**

The structure of the rectification network.

Layers	Out size	Configurations
Input	$1 \times 32 \times 100$	
Convolution	$64 \times 32 \times 100$	$3 \times 3, 64, s = 1, p = 1, b, r$
Maxpooling	$64 \times 16 \times 50$	$2 \times 2, s = 2$
Convolution	$128 \times 16 \times 50$	$3 \times 3, 128, s = 1, p = 1, b, r$
Maxpooling	$128 \times 8 \times 25$	$2 \times 2, s = 2$
Convolution	$64 \times 8 \times 25$	$3 \times 3, 64, s = 1, p = 1, b, r$
Convolution	$16 \times 8 \times 25$	$3 \times 3, 16, s = 1, p = 1, b, r$
Convolution	$2 \times 8 \times 25$	$3 \times 3, 2, s = 1, p = 1$
Maxpooling	$2 \times 8 \times 25$	$2 \times 2, s = 1, p = 1$
Tanh	$2 \times 8 \times 25$	

$n \times n$  denotes the size of kernels. 64, 128, 16 and 2 are channels of the output features.  $s, p$  are stride and padding sizes respectively.  $b, r$  denote the batch normalization layer and ReLU layer.

offset of the rectified pixel in the horizontal direction, and the second  $8 \times 25$  feature layer represents offset in the vertical direction. Each  $8 \times 25$  feature layer consists of float numbers in the range of  $[-1, 1]$ . Every  $4 \times 4$  pixels in the original image share the same offset, which can not only reduce the calculation but also make the rectified image smoother.

We adjust the input image based on  $\theta$ . Firstly, we resize  $\theta$  to the same size as the input image  $I$ , and obtain the offset map  $\theta'$  which corresponds to  $I$  point-to-point. Then, we normalize the coordinate of each pixel in  $I$  to get the initial grid  $g_1$ , whose coordinates are distributed in the interval  $[-1, 1]$ , where  $(-1, -1)$  represents the vertex in the upper left corner, and  $(1, 1)$  represents the vertex in the lower right corner.

Later on, we make two copies of  $g_1$  to represent the offsets in horizontal and vertical directions respectively, and add them with the corresponding channel of  $\theta'$  to get the offset mapping grid  $g_{offset}$ :

$$g_{offset} = g_c + \theta'_c \quad c = 1, 2 \quad (6)$$

We restore coordinates of  $g_{offset}$ 's two channels from  $[-1, 1]$  to  $[0, W]$  and  $[0, H]$  respectively. To get the pixel value of rectified image  $I'(i, j)$ , we first retrieve the position information  $(i', j')$  from the 0.1 channel at  $g_{offset}(i, j)$ , and then extract the pixel value from the  $(i', j')$  in the original image  $I$ .

$$(i', j') = g_{offset}(i, j) \quad (7)$$

$$I'(i, j) = I(i', j') \quad (8)$$

Compared with methods based on thin-plate-spline transformation [5], our method avoids the pressure of network initialization design in the training process, and is not subject to geometric constraints from the transformation matrix. By adjusting the pixel position in the horizontal and vertical directions, we can correct the irregular text image detailedly.

### 3.3. Recognition network

Our recognition network is a sequence-to-sequence recognition based on the attention mechanism. Only the input image and the corresponding text labels are needed. The structure of the recognition network is shown in Fig. 3, which is composed of the encoding part and the decoding part.

The role of the encoder is to extract discriminative features from the input image to improve the accuracy of the whole recognition model. Our encoder is a convolutional-recurrent neural network. Among several network structures proposed by He et al. [14], we choose 53-layers residual network to achieve a better trade-off between accuracy and calculation speed. It is also the choice of many other text recognition works such as Luo et al. [27], Shi et al. [35], Zhan and Lu [46], and it is conducive to a fair comparison of experimental results. Each residual unit is composed of  $1 \times 1$  and  $3 \times 3$  convolution operations. In order to retain more horizontal features, the first two residual units adopt the step size of  $2 \times 2$ , and then the convolution step size is adjusted to  $2 \times 1$ . After the residual network, we use two layers of bidirectional long short term memory units (BLSTM), each with 256 hidden layer units, to combine more context information and expand the receptive field of features obtained by ResNet.

The decoder is a sequence-to-sequence model, which transforms the feature sequence into a string sequence of any length. In [3,9] and [38], some different forms of sequence-to-sequence model are proposed. We adopt the GRU [8] model with 256 hidden units as our decoder. It works iteratively for  $T$  steps, and get a recognition result sequence  $(y_1, y_2, \dots, y_T)$  with the length of  $T$ .

In step  $t$ , the decoder calculates the current attention weight according to the output  $h$  of the encoder, the internal state  $s_{t-1}$  and the prediction result  $y_{t-1}$  of the last time. It can weigh the output  $h$  of the encoder, indicating the importance of the encoding features for the decoder. The attention mechanism is as follows:

$$a_{t,i} = \exp(e_{t,i}) / \sum_{i'=1}^L \exp(e_{t,i'}) \quad (9)$$

$$e_{t,i} = w^t \tanh(Ws_{t-1} + Vh_i + b) \quad (10)$$

where  $w, W, V$  are trainable weights and  $L$  is the length of the feature maps.

The glipse vector  $g_t$  is obtained by combining the current attention weight  $a_{t,i}$  with the output  $h$  of encoder. From the internal state  $s_{t-1}$  and the prediction result  $y_{t-1}$ , the current state vector  $s_t$  is calculated.

$$g_t = \sum_{i=1}^L (a_{t,i}, h_i) \quad (11)$$

$$s_t = \text{GRU}(f(y_{t-1}), g_t, s_{t-1}) \quad (12)$$

where  $f()$  represents the vector-encoding of  $y_{t-1}$ . The prediction result at time  $t$  is calculated by  $s_t$ , which can be a character or an end of sequence symbol (EOS):

$$y_t = \text{Softmax}(W_{\text{out}}s_t + b_{\text{out}}) \quad (13)$$

## 4. Experimental results

In this section, we conduct extensive experiments to verify the effectiveness of various aspects of ASTR and compare them with other classical and state-of-the-art methods.

### 4.1. Datasets

We evaluated the algorithm on seven public datasets: IIIT5K [28], SVT [20], IC03 [26], IC13 [22], IC15 [21], SVTP [30], and CUTE80 [31]. Among them, the first four datasets contain only regular text images, while the last three datasets contain irregular text images that include some degenerate factors such as radiation, bending, and blur. Some examples are shown in Fig. 1.

### 4.2. Implementatin details

The ASTR is implemented in Pytorch framework and runs on a server with 2.10 GHz CPU, GTX 1080Ti GPU, and Ubuntu 64-bit OS. We adopt ADADELTA [45] as the optimizer. In the training process, we first train the recognition network with learning rating 1. In the third stage of the curriculum learning strategy, we adjust the learning rate to 0.1 and join the rectification network for adversarial training. Another three stages later, the learning rate was adjusted to 0.01 to complete the training.

As described in Section 3, the specific structure of our rectification model and recognition model are described in Sections 3.2 and 3.3 respectively. We use synthetic data for training, including 8-million synthetic images released by Jaderberg et al. [16] and 6-million synthetic images released by Gupta et al. [13]. There is no additional training image and only word-level labels are needed, without character level annotation. In the testing process, instead of finetuning on the corresponding dataset, we directly test the images to get the recognition results.

### 4.3. Performance evaluation of the proposed method

#### 4.3.1. Ablation experiment

In order to verify the effectiveness of each module of ASTR, we designed ablation experiments to compare the recognition accuracy of each module on seven open datasets.

*The improvements to our baseline method* Our baseline method is a scene text recognition framework composed of rectification network and recognition network, in which the rectification module finds the control points at the upper and lower edge of the input image, and uses TPS transformation to get the rectified image. The recognition module is the same as our method, using the sequence to sequence recognition based on the attention mechanism. The whole network is trained in an end-to-end fashion. We use the “BaseLine\_I” to show the performance of the recognizer only (without rectification), indicating what proportion of data is actually suffering from degradation which requires rectification. “BaseLine\_II” is the full framework of the baseline method.

We verify the improvement of ASTR framework on the baseline by adding rectification network and adversarial loss separately. The overall recognition effect of ASTR framework on the test sets are shown in the last line. See Table 2 for specific experimental results:

From the results in Table 2, we can see that our adjustment of the rectification network is effective. At the same time, the proposed adversarial training strategy can improve the accuracy of the baseline method. The overall effect of ASTR reaches the optimal index of each dataset, which shows that the rectification module and recognition module we choose are more suitable for the proposed adversarial-learning strategy and achieve a better recognition effect.

*The improvements of adversarial learning to each part of ASTR*

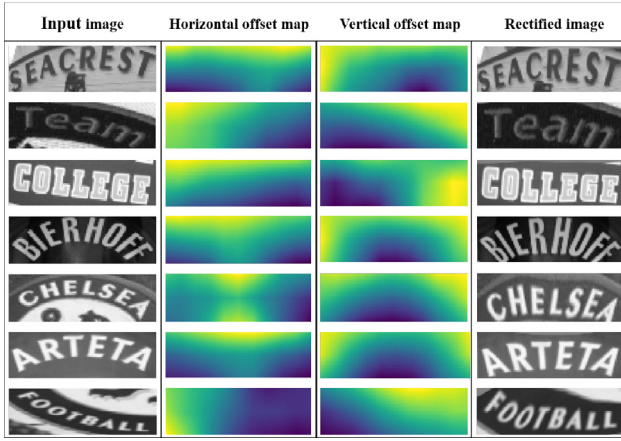


Fig. 4. Some examples of the rectification network.

**Table 2**  
Performance comparison of ablation experiments in different modules of ASTR.

Methods	IIIT5K	SVT	IC03	IC13	SVT-P	CUTE80	IC15
BaseLine_I	91	84.5	90.1	92.0	75.8	76.4	73.3
BaseLine_II	92.5	85.5	92.2	92.5	77.2	79.9	76.4
xy_Rectification	93	86.6	93.1	93.3	77.8	80.6	77.3
GANloss	92.9	86.2	92.3	93.7	79.8	82.2	78
ASTR	<b>93.5</b>	<b>89.6</b>	<b>93.6</b>	<b>94.3</b>	<b>81.2</b>	<b>82.3</b>	<b>80.2</b>

**Table 3**  
Performance comparison of recognizer under different training strategies.

	IIIT5K	SVT	IC03	IC13	SVT-P	CUTE80	IC15
end-to-end	91	84.5	90.1	92.0	75.8	76.4	73.3
GANloss	<b>92.8</b>	<b>86.4</b>	<b>93.1</b>	<b>93.1</b>	<b>78.8</b>	<b>80.6</b>	<b>77.4</b>

We hope to prove that our rectification network and recognition network can become more robust and perform better through adversarial learning. Therefore, we demonstrate the rectifying effect of the rectification network on irregular images in Fig. 4. It shows that after rectification, characters in the rectified image are more standard and less inclined, which is conducive to subsequent recognition.

**Table 4**  
Performance comparison of ASTR and other classical and STOA methods.

Method	IIIT5K [28]			SVT [20]		IC03 [26]			IC13 [22]	IC15 [21]	SVT-P [30]		CUTE80 [31]
	50	1K	0	50	0	50	Full	0			50	0	
Wang et al. [39]	–	–	–	57	–	76	62	–	–	–	–	–	–
Wang et al. [41]	–	–	–	70	–	90	84	–	–	–	–	–	–
Mishra et al. [29]	64.1	57.5	–	73.2	–	81.8	67.8	–	–	–	–	–	–
Jaderberg et al. [19]	–	–	–	86.1	–	96.2	91.5	–	–	–	–	–	–
Su and Lu [37]	–	–	–	83	–	92	82	–	–	–	–	–	–
Yao et al. [43]	80.2	69.3	–	75.9	–	88.5	80.3	–	–	–	–	–	–
Gordo [11]	93.3	86.6	–	91.8	–	–	–	–	–	–	–	–	–
Rodríguez-Serrano et al. [32]	76.1	57.4	–	70	–	–	–	–	–	–	–	–	–
Bissacco et al. [4]	–	–	–	–	–	90.4	78	–	87.6	–	–	–	–
Almazan et al. [1]	91.2	82.1	–	89.2	–	–	–	–	–	–	–	–	–
Jaderberg et al. [17]	97.1	92.7	–	95.4	80.7	98.7	<b>98.6</b>	93.1	90.8	–	–	–	–
Jaderberg et al. [15]	95.5	89.6	–	93.2	71.7	97.8	97	89.6	81.8	–	–	–	–
Shi et al. [33]	97.8	95	81.2	97.5	82.7	98.7	98	91.9	89.6	–	–	–	–
Shi et al. [34]	96.2	93.8	81.9	95.5	81.9	98.3	96.2	90.1	88.6	–	91.2	71.8	59.2
Lee et al. [23]	96.8	94.4	78.4	96.3	80.7	97.9	97	88.7	90	–	–	–	–
Yang et al. [42]	97.8	96.1	–	95.2	–	97.7	–	–	–	–	93	75.8	69.3
Cheng et al. [6]	99.3	97.5	87.4	97.1	85.9	<b>99.2</b>	97.3	94.2	93.3	66.2	92.6	71.5	63.9
Liu et al. [25]	97.7	94.5	83.3	95.5	83.6	96.9	95.3	89.9	89.1	–	94.3	73.5	–
Shi et al. [35]	–	–	92.67	–	–	–	–	93.72	90.74	–	–	78.76	76.39
Luo et al. [27]	97.9	96.2	91.2	96.6	88.3	98.7	97.8	<b>95</b>	92.4	68.8	94.3	76.1	77.4
ASTR	<b>99.4</b>	<b>98.6</b>	<b>93.5</b>	<b>97.8</b>	<b>89.6</b>	98.3	97.4	93.6	<b>94.3</b>	<b>81.2</b>	<b>96.9</b>	<b>80.2</b>	<b>82.3</b>

In Table 3, we compare the performance improvement of the recognizer by adversarial learning. In this experiment, we use the same structure of the rectification network and recognition network in ASTR. The difference is that the recognizer in the first row uses the traditional end-to-end learning strategy for training, while the other recognizer uses the adversarial learning strategy proposed by us. During test, we do not use the rectification network, but directly use the recognizer to recognize the original image. It can be seen clearly that the recognizer trained by the adversarial learning strategy has better recognition performance on each dataset. It further reflects the effectiveness of the proposed method.

#### 4.3.2. Comparison with other methods

In Table 4, we compare the recognition performance of ASTR with some classic and state-of-the-art methods. For the method of Shi [35], we use the experimental results corrected by the author in the open-source code. For the other methods, we use the experimental results provided by the author.

We can see that in the seven datasets mentioned in the first subsection, we achieve the best recognition performance on six datasets. Especially on some irregular datasets, ASTR has more obvious advantages. This further proves the robustness of our method and the effectiveness of ASTR in irregular scene text recognition task.

## 5. Conclusion

In this paper, the adversarial learning strategy is introduced into the scene text recognition task, and a new scene text recognition framework ASTR is proposed. In this algorithm, we regard the problem of irregular text rectification as a problem of regular text image generation and use the recognition network as a discriminator to evaluate the quality of the rectified image. At the same time, we use the difference between the original image and the rectified image in the recognition results to design the minimax game loss and achieve outstanding performance through adversarial learning.

The proposed method is validated on seven widely used scene text datasets. The experimental results show that our method has achieved state-of-the-art effect for the scene text recognition task. The outstanding performance on the irregular text datasets further

demonstrates the robustness and generality of the proposed training strategy.

For future works, we hope to explore the recognition network to directly extract the core features for text recognition on the bases of adversarial learning, so as to simplify the encoding and attention mechanism design process and further improve the recognition performance for irregular text images.

## Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant nos. 61531019, 71621002 and the Key Programs of the Chinese Academy of Sciences under Grant nos. ZDBS-SSW-JSC003, ZDBS-SSW-JSC004 and ZDBS-SSW-JSC005.

## References

- [1] J. Almazán, A. Gordo, A. Fornés, E. Valveny, Word spotting and recognition with embedded attributes, *IEEE Trans. Pattern Anal. Mach. Intell.* 36 (12) (2014) 2552–2566.
- [2] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S.J. Oh, H. Lee, What is wrong with scene text recognition model comparisons? dataset and model analysis, in: *International Conference on Computer Vision (ICCV)*, 2019. To appear
- [3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv:1409.0473* (2014).
- [4] A. Bissacco, M. Cummins, Y. Netzer, H. Neven, Photoocr: reading text in uncontrolled conditions, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 785–792.
- [5] F.L. Bookstein, Principal warps: thin-plate splines and the decomposition of deformations, *IEEE Trans. Pattern Anal. Mach. Intell.* 11 (6) (1989) 567–585.
- [6] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, S. Zhou, Focusing attention: towards accurate text recognition in natural images, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5076–5084.
- [7] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, S. Zhou, Aon: towards arbitrarily-oriented text recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5571–5579.
- [8] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
- [9] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y.N. Dauphin, Convolutional sequence to sequence learning, in: *Proceedings of the 34th International Conference on Machine Learning–Volume 70, JMLR.org*, 2017, pp. 1243–1252.
- [10] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.
- [11] A. Gordo, Supervised mid-level features for word image representation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2956–2964.
- [12] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: *Proceedings of the 23rd International Conference on Machine Learning*, ACM, 2006, pp. 369–376.
- [13] A. Gupta, A. Vedaldi, A. Zisserman, Synthetic data for text localisation in natural images, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2315–2324.
- [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [15] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Deep structured output learning for unconstrained text recognition, *arXiv preprint arXiv:1412.5903* (2014a).
- [16] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition, *Neural Information Processing Systems*, 2014.
- [17] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks, *Int. J. Comput. Vis.* 116 (1) (2016) 1–20.
- [18] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: *International Conference on Neural Information Processing Systems*, 2015, pp. 2017–2025.
- [19] M. Jaderberg, A. Vedaldi, A. Zisserman, Deep features for text spotting, in: *European Conference on Computer Vision*, Springer, 2014, pp. 512–528.
- [20] Kai Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: *2011 International Conference on Computer Vision*, 2011, pp. 1457–1464.
- [21] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V.R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, E. Valveny, Icdar 2015 competition on robust reading, in: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015, pp. 1156–1160.
- [22] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, G.I.B. Lluís, S. Robles Mestre, J. Mas, D. Fernandez Mota, J. Almazan Almazan, D.L.H. Lluís-Pere, Icdar 2013 robust reading competition, in: *International Conference on Document Analysis and Recognition*, 2013.
- [23] C.-Y. Lee, S. Osindero, Recursive recurrent nets with attention modeling for OCR in the wild, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2231–2239.
- [24] H. Li, P. Wang, C. Shen, G. Zhang, Show, attend and read: a simple and strong baseline for irregular text recognition, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 2019, pp. 8610–8617.
- [25] W. Liu, C. Chen, K.-Y.K. Wong, Z. Su, J. Han, Star-net: a spatial attention residue network for scene text recognition, in: *BMVC*, 2, 2016, p. 7.
- [26] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, Icdar 2003 robust reading competitions: entries, results, and future directions, *Int. J. Doc. Anal. Recognit.* 7 (2–3) (2005) 105–122.
- [27] C. Luo, L. Jin, Z. Sun, Moran: a multi-object rectified attention network for scene text recognition, *Pattern Recognit.* 90 (2019) 109–118.
- [28] A. Mishra, K. Alahari, C. Jawahar, Top-down and bottom-up cues for scene text recognition, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2687–2694.
- [29] A. Mishra, K. Alahari, C. Jawahar, Top-down and bottom-up cues for scene text recognition, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2687–2694.
- [30] T.Q. Phan, P. Shivakumara, S. Tian, C.L. Tan, Recognizing text with perspective distortion in natural scenes, in: *2013 IEEE International Conference on Computer Vision*, 2013, pp. 569–576.
- [31] A. Risnumawan, P. Shivakumara, C.S. Chan, C.L. Tan, A robust arbitrary text detection system for natural scene images, *Expert Syst. With Appl.* 41 (18) (2014) 8027–8048.
- [32] J.A. Rodriguez-Serrano, A. Gordo, F. Perronnin, Label embedding: a frugal baseline for text recognition, *Int. J. Comput. Vis.* 113 (3) (2015) 193–207.
- [33] B. Shi, X. Bai, 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.* 39 (11) (2016) 2298–2304.
- [34] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, Robust scene text recognition with automatic rectification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4168–4176.
- [35] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, Aster: an attentional scene text recognizer with flexible rectification, *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (2019) 2035–2048.
- [36] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, Scene text detection using graph model built upon maximally stable extremal regions, *Pattern Recognit. Lett.* 34 (2) (2013) 107–116.
- [37] B. Su, S. Lu, Accurate scene text recognition based on recurrent neural network, in: *Asian Conference on Computer Vision*, Springer, 2014, pp. 35–48.
- [38] I. Sutskever, O. Vinyals, Q. Le, Sequence to sequence learning with neural networks, *Adv. NIPS* 2 (2014) 3104–3112.
- [39] K. Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 1457–1464.
- [40] K. Wang, S. Belongie, Word spotting in the wild, in: *ECCV*, 2010, pp. 591–604.
- [41] T. Wang, D.J. Wu, A. Coates, A.Y. Ng, End-to-end text recognition with convolutional neural networks, in: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, IEEE, 2012, pp. 3304–3308.
- [42] X. Yang, D. He, Z. Zhou, D. Kifer, C.L. Giles, Learning to read irregular text with attention mechanisms, in: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 3280–3286.
- [43] C. Yao, X. Bai, B. Shi, W. Liu, Strokelets: a learned multi-scale representation for scene text recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4042–4049.
- [44] Q. Ye, D. Doermann, Text detection and recognition in imagery: a survey, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (7) (2014) 1480–1500.
- [45] M.D. Zeiler, Adadelta: an adaptive learning rate method, *arXiv preprint arXiv:1212.5701* (2012).
- [46] F. Zhan, S. Lu, Esir: end-to-end scene text recognition via iterative image rectification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2059–2068.
- [47] J. Zhao, C. Shi, F. Jia, Y. Wang, B. Xiao, Document image binarization with cascaded generators of conditional generative adversarial networks, *Pattern Recognit.* 96 (2019) 106968.
- [48] Y. Zhu, C. Yao, X. Bai, Scene text detection and recognition: recent advances and future trends, *Front. Comput. Sci.* 10 (1) (2016) 19–36.