An Attention-Based Recognizer for Scene Text

Yugang Li[1, *] and Haibo Sun[1]

Abstract: Scene text recognition (STR) is the task of recognizing character sequences in natural scenes. Although STR method has been greatly developed, the existing methods still can't recognize any shape of text, such as very rich curve text or rotating text in daily life, irregular scene text has complex layout in two-dimensional space, which is used to recognize scene text in the past Recently, some recognizers correct irregular text to regular text image with approximate 1D layout, or convert 2D image feature mapping to one-dimensional feature sequence. Although these methods have achieved good performance, their robustness and accuracy are limited due to the loss of spatial information in the process of two-dimensional to one-dimensional transformation. In this paper, we proposes a framework to directly convert the irregular text of two-dimensional layout into character sequence by using the relationship attention module to capture the correlation of feature mapping Through a large number of experiments on multiple common benchmarks, our method can effectively identify regular and irregular scene text, and is superior to the previous methods in accuracy.

Keywords: Scene text recognition, irregular text, attention.

1 Introduction

Scene text recognition (STR) is an import sub task in computer vision. Traffic sign reading, intelligent detection, image retrieval and many other practical applications all benefit from semantic information of scene text. With the development of scene text detection methods [Gómez and Karatzas (2017); Khare, Shivakumara, Raveendran et al. (2016)], scene character recognition has become the forefront of this research, which is considered as a very promising open and challenging research problem [Su and Lu (2017)].

STR solves the following problem: given an image patch that closely contains text from natural scenes (for example, license plates and posters on the street), what is the order of the characters? (1) At present, conventional text recognition methods [Bissacco, Cummins, Netzer et al. (2013); Shi, Bai and Yao (2016); Wang, Wu, Coates et al. (2012)] have achieved remarkable success. The application of deep neural network has greatly improved the performance of STR model [Sheng, Chen and Xu (2019); Lee and Osindero (2016)]. They usually combine the convolutional neural network (CNN) feature extractor used to extract input patches with the subsequent recurrent neural network (RNN) character sequence generator, which is responsible

[1] Academy of Broadcasting Science, Beijing, 100866, China.

[*] Corresponding Author: Yugang Li. Email: liyugang@abs.ac.cn.

Received: 18 February 2020; Accepted: 01 April 2020.

JAI. doi:10.32604/jai.2020.010203

for character decoding and language modeling. The model is trained in an end-to-end way, and the method based on convolutional neural network [Wang, Wu, Coates et al. (2012)] has been widely used. Combining recognition model with recurrent neural network [Shi, Wang, Lyu et al. (2016)] and attention mechanism [Lee and Osindero (2016); Yang, He, Zhou et al. (2017)] can get better performance.

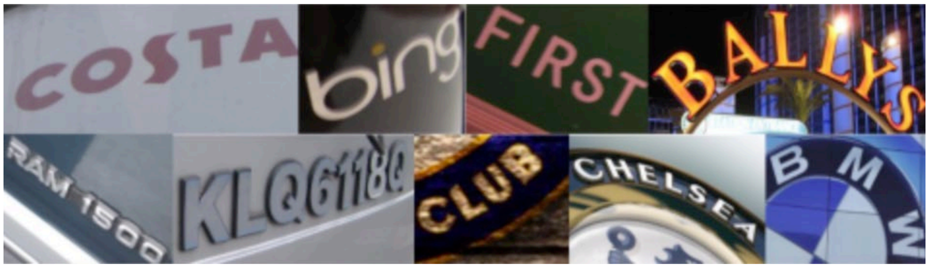

Figure 1: Examples of irregular scene text

However, most of the current recognition algorithms are unstable to deal with multiple interferences from the environment. In addition, various shapes and deformation patterns of text bring additional challenges. As shown in Fig. 1, text with various shapes, such as perspective text and curve text, is difficult to recognize. For example, Cheng et al. [Cheng, Bai, Xu et al. (2017)] and Shi et al. [Shi, Wang, Lyu et al. (2016); Shi, Yang, Wang et al. (2018)] have folded the height part of 2D CNN feature map into one-dimensional feature map, and they are unable to interpret arbitrarily shaped text conceptually and empirically, which is an important challenge in real deployment scenarios.

Recognizing the importance and difficulty of recognizing arbitrarily shaped text, the STR community attaches more importance to the introduction of "irregular shape" STR benchmark [Baek, Kim, Lee et al. (2019)] into such image types, which is the evidence of this interest. In terms of methods, recent STR methods pay more attention to the processing of irregular shaped text, mainly including two research routes: (1) input correction and (2) the use of input correction of two-dimensional feature map [Shi, Wang, Lyu et al. (2016); Shi, Yang, Wang et al. (2018)] the use of space converter network to standardize text image into a regular shape: horizontally aligned characters with consistent height and width. However, these methods are limited by the need to specify possible transformation families in advance.

On the other hand, some researchers use 2D feature map to obtain the original input image without any modification, then retrieve characters in 2D space in sequence [Yang, He, Zhou et al. (2017); Cheng, Xu, Bai et al. (2018); Li, Wang, Shen et al. (2019)]. Although the use of 2D feature map does add space for more complex modeling, the existing method's feature design is still limited by the assumption of input text level writing (SAR [Cheng, Xu, Bai et al. (2018)]), and the overly complex model Requirements for type a structure (AON [Cheng, Xu, Bai et al. (2018)]) or ground truth bounding box (ATR [Yang, He, Zhou et al. (2017)]). We believe that the community lacks a simple solution to deal with arbitrarily shaped text. This paper proposes a novel two-dimensional attention irregular scene text recognizer, as shown in Fig. 2. Different from the past, we use the two-dimensional attention module to modal the text information in the two-dimensional space through all pipelines. In order to achieve this goal, we first propose the relationship attention module to capture the global context information, rather than using RNN to model the context information, such as ATR [Yang, He, Zhou et al. (2017); Cheng, Xu, Bai et al. (2018); Li, Wang, Shen et al. (2019)]. In addition, we also designed a parallel attention module to generate masks on the two-dimensional feature map. By using the module, the proposed algorithm can output all characters at the same time, instead of predicting characters one by one like the previous methods [Yang, He, Zhou et al. (2017); Shi, Wang, Lyu et al. (2016); Li, Wang, Shen et al. (2019)]. Our model is an end-to-end framework, which does not need complex post-processing to detect and group characters and character level or pixel level annotations for training. The framework uses relation module to model local and global context information, and has strong robustness to complex irregular text (such as large curve text). Compared with the previous RNN scheme, the parallel module can predict all the results at the same time, so the method is more effective.

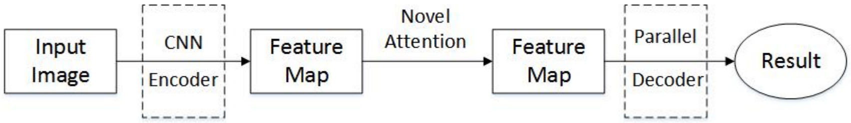

Figure 2: Irregular scene text recognizer

In order to verify the effectiveness of this method, we have carried out experiments on seven common benchmarks including regular and irregular data sets. We have obtained the latest results on almost all data sets, and proved the superiority of the algorithm. Our method is superior to the previous methods on SVTP and CUTE.

2 Related works

In academia, scene text recognition can be divided into regular text and irregular text. In this section, we will briefly review the relevant work in these two areas.

2.1 Scene text recognition on regular text

Regular text recognition is the focus of early research. Mishra et al. [Mishra, Alahari and Jawahar (2016)] used the traditional sliding window based method to describe the bottom-up prompts, and used the glossary before modeling the top-down prompts. The combination of these two clues is to minimize the energy of character combination. Shi et al. [Shi, Bai and Yao (2016)] proposed an end-to-end trainable character free annotation network, called CRNN. CRNN uses CNN to extract one-dimensional feature sequence, and then uses RNN to encode the sequence, and finally calculates CTC loss. This is the first work that only needs word level annotation but not character level annotation. Gao et al. [Gao, Chen, Wang et al. (2019)] integrated the attention module into the remaining blocks, amplifying the foreground response and suppressing the background response. However, the attention module cannot encode the global dependency among pixels Human [Cheng, Bai, Xu et al. (2017)] observed that attention may drift due to complex scenes or low-quality images, which is a weakness of ordinary two-dimensional attention network. In order to solve the problem of misalignment between the input sequence and the target, Bai et al. [Bai, Cheng, Niu et al. (2018)] adopted the attention based encoder decoder architecture, and estimated the editing probability of the text based on the output sequence. Editing probability is aimed at the problem of character loss and redundancy. Zhang et al. [Zhang, Nie, Liu et al. (2019)] the unsupervised fixed length do main adaptive method is used for the variable length scene text recognition region, and the model is also based on the attention coding decoding structure.

2.2 Scene text recognition on irregular text

Irregular text recognition is more challenging than conventional text recognition. However, it has attracted the efforts of most researchers. Yao et al. [Yao, Bai, Liu et al. (2012)] is one of the first works that explicitly put forward the multi-directional text detection model. Shi et al. [Shi, Yang, Wang et al. (2018)] attempts to solve the problem of multi-type irregular text recognition within a framework through spatial transformation network (STN) [Jaderberg, Simonyan and Zisserman (2015)]. In order to further improve recognition performance, Zhan and Lu [Zhan and Lu (2019)] proposed forward parallel iterative correction for text image. In this method, the attitude of the text line is estimated by learning the middle line of the text line and L line segment needed by the thin plate spline. However, the rectification-based method is constrained to -10 by character geometry, and the background noise will be magnified unexpectedly. In order to overcome this problem, Luo et al. [Luo, Jin and Sun (2019)] proposed a multi-target corrected attention network which is more flexible than the direct affine transform estimation. Unlike the rectification-based method, show attention read (SAR) proposed by Li et al. [Li, Wang, Shen et al. (2019)] uses 2D attention mechanism to guide the encoder decoder recognition module to focus on the corresponding character region. This method does not need complex space transformation.

2D attention can represent the relationship between target output and input image features, while ignoring the global context between pixels and the potential correlation between characters in [Hu, Gu, Zhang et al. (2018)]. An object relationship module is proposed. After the success of transformer [Vaswani, Shazeer, Parmar et al. (2017)], Wang et al. [Wang, Girshick, Gupta et al. (2018)] added a self focus block to the non local network. Recently, Sheng et al. [Sheng, Chen and Xu (2019)] proposed a scene text recognizer based on transformer, which can learn the self focus of encoder and decoder. It uses a simple CNN module to extract 1D sequence features and input them into a transformer to decode the target output. However, the transformer's self focus module consists of several fully connected layers, which greatly increases the number of parameters. Wang et al. [Wang, Girshick, Gupta et al. (2018)] gave up the encoder of the original transformer and only kept CNN feature extractor and decoder for irregular scene text recognition. However, it can't encode the global context of pixels in feature mapping. The network proposed in this paper not only learns the two-dimensional attention between the input feature and the output target, but also learns the self-attention inside the feature extractor and decoder. The non-local blocks can encode different types of spatial feature dependencies with low computational cost and compact model.

In this paper, we propose a text recognizer for scene text with 2D attention, which can directly convert the two-dimensional layout of irregular text into character sequence.

Compared with the rectification pipeline, this method is more robust and effective for text image recognition in irregular scene. In addition, compared with the previous method in two-dimensional perspective, this method has parallel attention mechanism and doesn't need character-level localization annotations.

3 Methodology

The proposed algorithm structure is shown in Fig. 3. We first use CNN encoder to encode the input image with high-level semantic information. Then, the relational attention module is applied to each pixel of feature mapping to obtain the global correlation. Then, based on the output of relational attention module, a parallel attention module is constructed and a certain amount of flash is output. Finally, the character decoder decodes these flashes into characters.

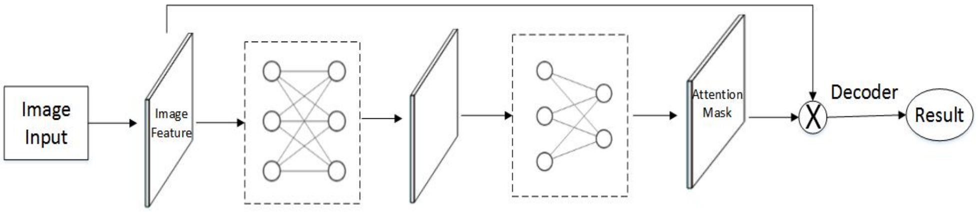

Figure 3: Network structure

3.1 Attention module

Inspired by Vaswani et al. [Vaswani, Shazeer, Parmar et al. (2017)], we capture the global dependency between input and output by aggregating information from input elements, and construct a new attention module composed of converter elements proposed in Vaswani et al. [Vaswani, Shazeer, Parmar et al. (2017)]. The attention module captures global information in parallel, which is more effective than the above strategies. Specifically, following BERT, our attention module architecture is a multi-layer bidirectional transformer encoder.

The attention module is composed of transformer units. For the first layer, the input is the sum of input image feature and position embedding feature. For other layers, the input is the output of the previous transformer layer.

By flatting input feature into a k×c sequence, we can process arbitrary shape input with the attention module. Where k is the length and c is the feature dimensions. For each feature, we construct a position vector which has the same dimension with feature. Then, we fuse the constructed position vector with the previous processed input features to obtain the position sensitive fusion features.

The attention module consists of multiple transformer layers. We think this structure can obtain better aggregate information from the input. Finally, we using the last layer's output as next module's input.

The basic attention module used in Yang et al. [Yang, He, Zhou et al. (2017); Shi, Wang, Lyu et al. (2016); Shi, Yang, Wang et al. (2018); Li, Wang, Shen et al. (2019)] and integrated with a RNNs unit:

$$\alpha_t = Att(h_{t-1}, \alpha_{t-1}, I) \tag{1}$$

where h_{t-1} and α_{t-1} are the hidden state and attention weights of the RNN decoder at the previous step, *I* means the encoded image feature sequence. As formulated in Eq. (1), the computation of the step *t* is limited by the previous steps, which is inefficient.

We propose a parallel attention module, which does not need repeated attention, but eliminates the dependency between output nodes. Therefore, for each output node, the calculation of attention is independent and can be easily implemented and optimized in parallel.

Specifically, we assign the number of output nodes to *n*. Given a feature sequence *E* in the shape of $k \times c$, the parallel attention module outputs the weight coefficient α as formulated in Eq. (2).

$$\alpha = soft\max(W_2 \tanh(W_1 E^T)) \tag{2}$$

Here, W_1 and W_2 are the learnable parameters with the shape of $c \times c$ and $n \times c$ respectively.

Based on the weight coefficients α and the encoded image feature sequence *I*, the glimpses of each output node can be obtained by Eq. (3).

$$outputG_i = \sum_{j=1}^{k} \alpha_{ij} I_j \tag{3}$$

where *i* and *j* are the index of output node and feature vector respectively.

3.2 Decoder

Although the above attention module is more effective than the basic attention model, due to the parallel independent computing, the dependency between output nodes will be lost. In order to capture the dependencies of output nodes, we build a parallel decoder.

The first layer uses character decoder directly to predict characters. The second level decoder obtains a "glimpse" through a traditional relational attention module, then sends it to the character decoder for prediction.

We optimize the network in an end-to-end way. Because the decoder is a parallel structure, we directly optimize it as a multi-task loss. The loss function is formulated in Eq. (4).

$$loss = \sum_{i=1}^{2} \sum_{j=1}^{n} (-y_j^{'} \log y_j) \tag{4}$$

where $y_j^{'}$ is the real text sequence, y_j is the *j*-th character of prediction, *i* and *j* are the index of decoder and output node.

4 Experiments

We have carried out extensive experiments on several benchmarks to verify the effectiveness of our method, and compare it with the state-of-the-art methods. In all the experiments in this paper, we use the accuracy as the evaluation criteria. That is, the proportion of the number of characters recognized to the total number of characters recognized.

4.1 Datasets

In this paper, we conduct experiments on a number of datasets to verify the effectiveness of our proposed model. The model is trained on three synthetic datasets: Synth90K, Synth-Text and SynthAdd, and evaluated on both regular and irregular scene text datasets.

The training datasets consist of the following datasets.

Synth90k (MJSynth) is the synthetic text dataset. The dataset has 9 million images generated from a set of 90k common English words. Every image in Synth90k is annotated with a word-level ground-truth. All of the images in this dataset are used for training.

SynthText is a synthetic text dataset originally introduced for text detection. The generating procedure is similar to Synth90k, but different from Synth90k, words are rendered onto a full image with large resolution instead of a text line. 800 thousand full images are used as background images, and usually each rendered image contains around 10 text lines. Recently, it is also widely used for scene text recognition. We obtain 7 millions of text lines from this dataset for training.

SynthAdd is the synthetic text dataset. The dataset contains 1.6 million word images using the synthetic engine proposed by Synth90k to compensate the lack of special characters like punctuations. All of the images in this dataset are used for training.

The test datasets consist of the following datasets.

ICDAR 2003 (IC03), ICDAR 2013(IC13), Street View Text (SVT) and IIIT5k-Words (IIIT5K) are regular text dataset. Their images are cropped from real scene text images or Google Street View images. Before using datasets, we removed some samples which contain non-alphanumeric characters or have less than three characters. Meanwhile, we choose 50 words lexicon for each sample.

ICDAR 2015 Incidental Text (IC15), SVT-Perspective (SVTP) and CUTE80 (CUTE) are irregular text dataset. The datasets are respectively from images taken by Google glasses, street view images and specific curve text. They are usually used to evaluate the performance of algorithms in recognizing multi-angle text, perspective text and curve text.

4.2 Implementation details

Our model is implemented using PyTorch. We use two NVIDIA TITAN XP GPU parallel training the model, then evaluate it on a single GPU with the batch size of 1.

We train the model from scratch using Adam optimizer and cross-entropy loss with a batch size of 512. The learning rate is set to be 4×10^{-4} over the whole training phase. We observe that the learning rate should be associated with the number of GPUs. For one GPU, 1×10^{-4} is a good choice. Our model is trained for 12 epochs, each epoch takes about 3 hours.

In the test stage, for the image whose height is greater than the width, we rotate the image 90 degrees clockwise and counter clockwise to feed the original image and two rotated images into the model, and select the output result with the maximum output probability.

4.3 Comparisons with state-of-the-arts

As shown in the Tab. 1, we conducted relevant experiments on two different types of test datasets. In regular scene setting, the result of our algorithm is better than the previous models on most datasets. For IIIT5K and SVT with some curved or text oriented examples, our method obtains 92.1% and 90.4% performance and are better than most proposed methods. Especially in IC03 and IC13, we obtain the state-of-the-art results which are 95.6% and 93.5%. Meanwhile, the proposed method has better performance than [Cheng, Bai, Xu et al. (2017)] which is designed for regular text recognition task. These experiments results show that our algorithm has certain generality and robustness in regular text recognition.

Table 1: Results on the public datasets

Method	Regular test dataset				Irregular test dataset		
	IIIT5K	SVT	IC03	IC13	IC15	SVTP	CUTE
Shi et al. [Shi, Wang, Lyu et al. (2016)]	81.9	81.9	-	-	-	71.8	59.2
Wang and Hu 2017	80.8	81.5	-	-	-	-	-
Cheng et al. [Cheng, Bai, Xu et al. (2017)]	87.4	85.9	94.2	93.3	-	-	-
Shi et al. [Shi, Yang, Wang et al. (2018)]	93.4	93.6	-	91.8	76.1	78.5	79.5
Baek et al. [Baek, Kim, Lee et al. (2019)]	87.9	87.5	94.4	92.3	71.8	79.2	79.9
Zhan et al. [Zhan and Lu (2019)]	93.3	90.2	-	-	76.9	79.6	83.3
Yang et al. [Yang, He, Zhou et al. (2017)]	-	-	-	-	-	75.8	69.3
Cheng et al. [Cheng, Xu, Bai et al. (2018)]	87.0	82.8	91.5	91.5	68.2	73.0	76.8
Li et al. [Li, Wang, Shen et al. (2019)]	91.5	84.5	-	-	69.2	76.4	83.3
Ours	92.1	90.4	95.6	93.5	77.8	80.3	84.7

We use three test datasets of irregular text to evaluate the performance of our model. In IC15, SVTP and CUTE, our method obtains 77.8%, 80.3% and 84.7 performance and is 0.9%, 0.7% and 1.4% better than the previous the state-of-the-art results respectively. In detail, our model's performance surpasses the rectification-based method [Shi, Yang, Wang et al. (2018)] above 5.2% and performs better than [Cheng, Xu, Bai et al. (2018)] which recognizes irregular text by 7.9%. From the experimental results, we can find that our method has better performance in the recognition of irregular scene text, especially in the recognition of more complex text image, with stronger robustness and effectiveness.

5 Conclusion

In this paper, a new text recognition method for irregular scenes is proposed. The text image recognition method based on two-dimensional image features can directly retain and utilize the two-dimensional spatial information of text. In addition, due to the effectiveness of the proposed attention module, our method is better than the previous one. We evaluated the performance of our model on seven public datasets and proved the effectiveness of the method.

In future, it is worth extending this method to deal with arbitrary-oriented text recognition, which is more challenging due to the wide variety of text and background. Moreover, recognizing text in vertical and more complex layouts will be a goal for us.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D. et al. (2019): What is wrong with scene text recognition model comparisons? dataset and model analysis. *IEEE International Conference on Computer Vision*, pp. 4715-4723.

Bai, F.; Cheng, Z.; Niu, Y.; Pu, S.; Zhou, S. (2018): Edit probability for scene text recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1508-1516.

Bissacco, A.; Cummins, M.; Netzer, Y.; Neven, H. (2013): Photoocr: Reading text in uncontrolled conditions. *IEEE International Conference on Computer Vision*, pp. 785-792.

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S. et al. (2017): Focusing attention: Towards accurate text recognition in natural images. *IEEE International Conference on Computer Vision*, pp. 5076-5084.

Cheng, Z.; Xu, Y.; Bai, F.; Niu, Y.; Pu, S. et al. (2018): Aon: Towards arbitrarily-oriented text recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5571-5579.

Gao, Y.; Chen, Y.; Wang, J.; Tang, M.; Lu, H. (2019): Reading scene text with fully convolutional sequence modeling. *Neurocomputing*, vol. 339, pp. 161-170.

Gómez, L.; Karatzas, D. (2017): Textproposals: a text-specific selective search algorithm for word spotting in the wild. *Pattern Recognition*, vol. 70, pp. 60-74.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. (2018): Relation networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588-3597.

Jaderberg, M.; Simonyan, K.; Zisserman, A. (2015): Spatial transformer networks. *Advances in Neural Information Processing Systems*, pp. 2017-2025.

Khare, V.; Shivakumara, P.; Raveendran, P.; Blumenstein, M. (2016): A blind deconvolution model for scene text detection and recognition in video. *Pattern Recognition*, vol. 54, pp. 128-148.

Lee, C. Y.; Osindero, S. (2016): Recursive recurrent nets with attention modeling for ocr in the wild. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2231-2239.

Li, H.; Wang, P.; Shen, C.; Zhang, G. (2019): Show, attend and read: A simple and strong baseline for irregular text recognition. *AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8610-8617.

Luo, C.; Jin, L.; Sun, Z. (2019): Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, vol. 90, pp. 109-118.

Mishra, A.; Alahari, K.; Jawahar, C. V. (2016**):** Enhancing energy minimization framework for scene text recognition with top-down cues. *Computer Vision and Image Understanding*, vol. 145, pp. 30-42.

Sheng, F.; Chen, Z.; Xu, B. (2019): NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. *International Conference on Document Analysis and Recognition*, pp. 781-786.

Shi, B.; Bai, X.; Yao, C. (2016): 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*, vol. 39, no. 11, pp. 2298-2304.

Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. (2016): Robust scene text recognition with automatic rectification. *IEEE Conference on CVPR*, pp. 4168-4176.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C. et al. (2018): Aster: An attentional scene text recognizer with flexible rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2035-2048.

Su, B.; Lu, S. (2017): Accurate recognition of words in scenes without character segmentation using recurrent neural network. *Pattern Recognition*, vol. 63, pp. 397-405.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L. et al. (2017): Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998-6008.

Wang, T.; Wu, D. J.; Coates, A.; Ng, A. Y. (2012): End-to-end text recognition with convolutional neural networks. *International Conference on Pattern Recognition*, pp. 3304-3308.

Wang, X.; Girshick, R.; Gupta, A.; He, K. (2018): Non-local neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794-7803.

Yang, X.; He, D.; Zhou, Z.; Kifer, D.; Giles, C. L. (2017): Learning to read irregular text with attention mechanisms. *International Joint Conference on Artificial Intelligence*, pp. 3280-3286.

Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. (2012**):** Detecting texts of arbitrary orientations in natural images. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1083-1090.

Zhan, F.; Lu, S. (2019): ESIR: End-to-end scene text recognition via iterative image rectification. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2059-2068.

Zhang, Y.; Nie, S.; Liu, W.; Xu, X.; Zhang, D. et al. (2019): Sequence-to-sequence domain adaptation network for robust text image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2740-2749.