STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition

Wei Liu[1]
wliu@cs.hku.hk
Chaofeng Chen[1]
cfchen@cs.hku.hk
Kwan-Yee K. Wong[1]
kykwong@cs.hku.hk
Zhizhong Su[2]
suzhizhong@baidu.com
Junyu Han[2]
hanjunyu@baidu.com

[1] Department of Computer Science
The University of Hong Kong, HK

[2] Institution of Deep Learning
Baidu Inc, Beijing

Abstract

In this paper, we present a novel SpaTial Attention Residue Network (STAR-Net) for recognising scene texts. Our STAR-Net is equipped with a spatial attention mechanism which employs a spatial transformer to remove the distortions of texts in natural images. This allows the subsequent feature extractor to focus on the rectified text region without being sidetracked by the distortions. Our STAR-Net also exploits residue convolutional blocks to build a very deep feature extractor, which is essential to the successful extraction of discriminative text features for this fine grained recognition task. Combining the spatial attention mechanism with the residue convolutional blocks, our STAR-Net is the deepest end-to-end trainable neural network for scene text recognition. Experiments have been conducted on five public benchmark datasets. Experimental results show that our STAR-Net can achieve a performance comparable to state-of-the-art methods for scene texts with little distortions, and outperform these methods for scene texts with considerable distortions.

1 Introduction

Scene text recognition refers to recognising words that appear in various kinds of natural images. It has received much attention as many real world applications can benefit from the rich semantic information embedded in the scene text images. Examples include self-driving cars, text-to-speech devices for visually-impaired people, and image-based machine translation softwares. In recent years, remarkable progress in scene text recognition has been achieved for tightly bounded text images with no severe distortion [15, 19, 20, 38]. In general scenarios, however, scene texts often do not occupy the entire image, and they suffer from various kinds of distortions (see Figure 1(a)). Hence, it remains an open challenge to build a recogniser that is capable of handling such loosely bounded and distorted scene texts.

(a) (b)

Figure 1: Text images from some public benchmark datasets and images rectified by our spatial attention module. (a) At the top are four tightly bounded, horizontal and frontal text images; at the bottom are text images suffering from different kinds of distortions. (b) Text images with different kinds of distortions are rectified by our spatial attention mechanism.

The literature is relatively sparse when it comes to the problem of handling spatial distortion in text recognition. Phan *et al*. [30] employed the SIFT descriptor as an invariant feature for recognising perspective scene texts. Instead of using hand-crafted features, we tackle this problem using a spatial attention mechanism that is capable of directly outputting the transformation parameters required for the rectification. Figure 1(b) shows some rectification results using our spatial attention mechanism. This spatial attention mechanism transforms a distorted text region into a canonical pose suitable for recognition. This greatly eases the difficulties in recognising loosely bounded and severely distorted texts. During the course of preparation of this paper, Shi *et al*. [33] published a similar idea for handling distorted text regions. Similar to the work by Lee and Osindero [24], they introduced a RNN-based attention model from neural machine translation [7, 35]. Their RNN-based attention model focused on how to translate the features extracted from an input image into the corresponding sequence of labels. In scene text recognition, this can be considered as a problem of image-based sequence-to-sequence classification. Instead of using RNN-based attention models and focusing on how to translate the extracted text features, we emphasise the importance of representative image-based feature extraction from text regions using a spatial attention mechanism and a residue learning strategy [14].

In this paper, we propose a novel deep residue network with a spatial attention mechanism for unconstrained scene text recognition. We name our network **SpaTial Attention Residue Network (STAR-Net)**. On one hand, we adopt a spatial transformer module [21] to introduce spatial attention into our network. This guarantees the full potential of the subsequent feature extractor is exploited in extracting discriminative text features, rather than being tolerant to spatial distortions. On the other hand, because of the recent success of residue learning in image classification [14, 36], demonstrating its strong capability in learning representative image features, we adopt residue convolutional blocks to build a feature extractor with more convolutional layers than [24, 32, 33]. Such a deep feature extractor can extract discriminative text features suitable for this fine grained recognition task. The key contributions of this work are

1. An end-to-end trainable STAR-Net which is a novel deep neural network integrating spatial attention mechanism and residue learning for scene text recognition. Experimental results show that it outperforms other state-of-the-art methods for loosely bounded text images with considerable distortions.

2. A spatial attention mechanism that can simultaneously locate the text region and eliminate its geometrical distortion without the need of any direct supervision. This spatial attention mechanism transforms a distorted text region into a canonical pose suitable for recognition. This greatly eases the difficulties in recognising loosely bounded and severely distorted texts, and allows the subsequent feature extractor to fully devote its power to extracting discriminative text features, rather than being tolerant to spatial distortions.

3. A deep feature extractor builds upon residue convolutional blocks, with the extremely deep convolutional layers and one recurrent layer (BLSTM) [11] being optimised under the non-parameterised supervision of Connectionist Temporal Classification (CTC) [13]. Note that this is the very first time that residue convolutional blocks demonstrate strong capabilities for scene text recognition.

2 Methodology

In the following sections, we describe in detail the three key components of our STAR-Net, namely the spatial transformer (Section 2.1), the residue feature extractor (Section 2.2) and the connectionist temporal classification (Section 2.3). Figure 2 illustrates the overall architecture of our STAR-Net.

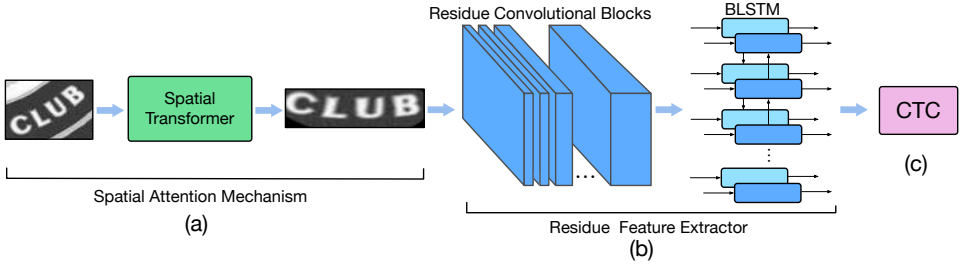


Figure 2: Overview of our STAR-Net for scene text recognition. (a) Spatial attention mechanism. (b) Residue feature extractor. Residue convolutional blocks with different widths represent different sizes of feature channels. (c) Connectionist Temporal Classification.

2.1 Spatial Transformer

The spatial transformer is responsible for introducing the spatial attention mechanism, by transforming a loosely bounded and distorted text region into a more tightly bounded and rectified text region. It enables the subsequent feature extractor to fully focus on extracting

discriminative text features instead of being tolerant to spatial distortions. The spatial transformer is composed of three parts, namely localisation network, sampler and interpolator (see Figure 3(a)). The localisation network is used to determine the distortion exhibited by the original text image and outputs the corresponding transformation parameters. Based on these parameters, the sampler locates sampling points on the input image which explicitly define the text region to be aed. Finally, the interpolator generates the output image by interpolating the intensity values of the four pixels nearest to each sampling point. In order to simplify the explanation, we use a spatial transformer with an affine transformation to illustrate the idea.

Localisation Network The localisation network takes the original grey image $\mathbf{I} \in \mathbb{R}^{W \times H}$ with width W and height H as input, and directly outputs the parameters of an affine transformation

$$\theta(\mathbf{I}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}. \tag{1}$$

In this work, the localisation network $\theta(\cdot)$ takes the form of a convolutional neural network, which includes a final regression layer to predict all the transformation parameters. We do not have any direct supervision towards the transformation parameters. The network will be trained to learn the suitable parameters for different inputs according to the gradient back-propagated from the recognition objective function (Eq. 7).

Sampler The sampler aims at locating a sampling point on the input image for every pixel in the output image $\mathbf{I}' \in \mathbb{R}^{W' \times H'}$ (i.e., the rectified image). A sampling point (x_i, y_i) on the input image for a pixel (x'_i, y'_i) on the output image can then be computed using the parameters $\theta(\mathbf{I})$ in Eq. 1 as follows,

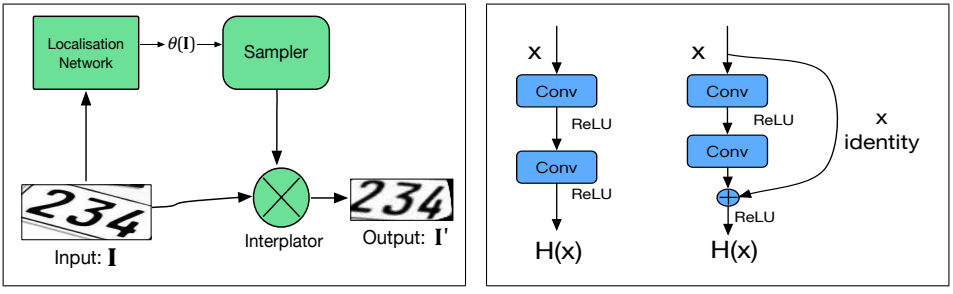
$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = \theta(\mathbf{I}) \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix}. \tag{2}$$

Interpolator The interpolator generates an intensity value for a pixel (x'_i, y'_i) in the output image from the intensity values of the four pixels in the input image which are nearest to the sampling point (x_i, y_i). In this work, we employ bilinear interpolation to compute the intensity values from those of the four nearest pixels.

Note that all the equations are differentiable. This allows the spatial transformer to be optimised easily using a gradient descent algorithm.

2.2 Residue Feature Extractor

To fully exploit the potential of convolutional layers and build up a powerful deep feature encoder, we employ residue convolutional blocks for extracting image-based features, and use Long Short-Term Memory (LSTM) [16] for encoding long-range dependencies among sequential features. Note that the feature map outputted from the convolutional neural network is in the spatial domain, and has a dimension of $C_s \times H_s \times W_s$, where C_s, H_s and W_s denote the channels, height and width of the feature map respectively. We transform this three-dimensional feature map into sequential features by cutting it along its width into W_s 2D slices, each with a dimension of $C_s \times H_s$, and remapping each slice into a vector $\vec{s}_t$, where $t = [1, 2, \ldots, W_s]$.



(a) (b)

Figure 3: Structures of the spatial transformer, plain and residue convolutional blocks. (a) Spatial transformer; (b) On the left is the plain convolution block; on the right is the residue convolutional block.

Residue Convolutional Block The residue block used in our encoder includes two convolutional layers, two ReLU[28] activations and a shortcut connection between the input and the output of the second convolutional layer (see Figure 3(b)). Let x denote the input to the block, and $H(x)$ is a complicated function that we want to approximate. A plain convolutional block targets at finding suitable parameters ($\mathbf{W}_{\text{pb}}$) of the convolutional layers to approximate $H(x)$ directly, *i.e.*,

$$H(x) = PB(x, \mathbf{W}_{\text{pb}}), \tag{3}$$

whereas a residue convolutional block, with the special shortcut connection, targets at finding suitable parameters ($\mathbf{W}_{\text{rb}}$) of the convolutional layers to approximate the residue function $H(x) - x$, *i.e.*,

$$H(x) = RB(x, \mathbf{W}_{\text{rb}}) + x. \tag{4}$$

Although both blocks should be capable of approximating $H(x)$, the ease of optimisation is different. In order to avoid the degradation problem[1] caused by adding more plain convolutional blocks, STAR-Net employs the residue convolutional block as a basic component to build an extremely deep feature extractor with 18 convolutional layers.

Long Short-Term Memory Long Short-Term Memory (LSTM) is a kind of recurrent layer capable of learning long-range dependencies of the input sequential features. The basic component of LSTM is the memory block. Each memory block consists of a memory cell c, an input gate i, a forget gate f and an output gate o, respectively, and can be formulated as

$$\begin{aligned}
\vec{i}_t &= \sigma(\mathbf{W}_i \times [\vec{h}_{t-1}, \vec{s}_t] + \vec{b}_i), \quad g \in \{i, f, o\} \\
\vec{f}_t &= \sigma(\mathbf{W}_f \times [\vec{h}_{t-1}, \vec{s}_t] + \vec{b}_f), \\
\vec{o}_t &= \sigma(\mathbf{W}_o \times [\vec{h}_{t-1}, \vec{s}_t] + \vec{b}_o), \\
\vec{c}_t &= \vec{f}_t \cdot \vec{c}_{t-1} + \vec{i}_t \cdot \tanh(\mathbf{W}_c \times [\vec{h}_{t-1}, \vec{s}_t] + \vec{b}_c), \\
\vec{h}_t &= \vec{o}_t \cdot \tanh(\vec{c}_t),
\end{aligned} \tag{5}$$

[1] He *et al.* [14] point out that with the network depth increasing, accuracy gets saturated and degrades rapidly.

where the $\mathbf{W}$ terms denote the weights (*e.g.*, $\mathbf{W}_i$ is the weight of the input gate), $\vec{b}$ denote the bias (*e.g.*, $\vec{b}_i$ is the bias of the input gate), $\vec{h}$ denote the hidden states, and σ is the logistic sigmoid function. In this work, we use one Bidirectional-LSTM layer [11] which calculates $\vec{h}_t$ (where $t = 1, 2, \ldots, W_s$), using the memory block from both forward and backward directions (see the BLSTM part in Figure 2).

2.3 Connectionist Temporal Classification

One of the distinctive properties of CTC is that there are no parameters to be learned for decoding. This addresses our goal in emphasising the importance of our feature extractor on extracting discriminative sequential features by keeping our decoder simple. Let L denote the set of 36-class (26 letters and 10 digits) case insensitive characters in our task and $L' = L \cup \{blank\}$. At the last step of our encoder, we adopt a softmax layer to output a probability map $\mathbf{y} = \{\vec{y}^1, \ldots, \vec{y}^{W_s}\}$ conditioned on the sequential features $\vec{s}_t$. Each $\vec{y}^{\,t}$ is a probability distribution over L', and $\vec{y}^{\,t}_m$ represents the probability of observing the label m at time t. The probability of a given sequence π with length W_s is defined as $p(\pi) = \prod_{t=1}^{W_s} \vec{y}^{\,t}_{\pi_t}$. As in [13], we define the map-to-one map $\mathcal{B}$ that removes the repeated labels and then all the blanks from the output sequence (*e.g.*, $\mathcal{B}(--a-bb-d-c) = abdc$). For a given labelling $\mathbf{l}$ with length T ($T \leq W_s$), its probability can be formulated as

$$p(\mathbf{l}|\mathbf{y}) = \sum_{\pi} p(\pi),\ \pi \in \mathcal{B}^{-1}(\mathbf{l}), \tag{6}$$

where $\mathcal{B}^{-1}(\mathbf{l})$ denotes the set of sequences with length W_s which are mapped to $\mathbf{l}$ by $\mathcal{B}$. The objective function O_{ctc} of CTC is defined as the sum of the negative log likelihood of Eq. 6 over the whole training set $\mathbb{S}$:

$$O_{ctc} = -\sum_{(\mathbf{l},\mathbf{y})\in\mathbb{S}} \ln p(\mathbf{l}|\mathbf{y}). \tag{7}$$

Since the number of sequences corresponding to a given labelling increases exponentially with W_s, a dynamic programming named Forward-Backward algorithm (see [12]) is used for calculating O_{ctc} and its partial derivatives $\frac{\partial O_{ctc}}{\partial \vec{y}^{\,t}_{\pi_t}}$ efficiently when training the final CTC layer. In the phase of predicting the sequence of labels, we simply pick the label with the highest probability at each time step and then apply $\mathcal{B}$ to the entire output path.

3 Experiment

In this section, we evaluate the performance of our STAR-Net for scene text recognition on the following five public benchmarks.

- **ICDAR-2003** [26] contains 251 full-size scene images and 860 cropped text images for testing. Each cropped text image has a 50-word lexicon defined by Wang *et al.* [37]. A full lexicon is constructed with all 50-word lexicons. Following the protocol proposed by Wang *et al.* [37], we recognise the images containing only alphanumeric words (0-9 and A-Z) with at least three characters.

- **ICDAR-2013** [23] is derived from ICDAR-2003, and contains 1,015 cropped word test images without any pre-defined lexicon.

- **IIIT5K** [27] contains 3,000 cropped text images for testing. These images are collected from the Internet and each image has a 50-word lexicon and a 1,000-word lexicon.

- **Street View Text** [37] contains 647 test word images which are cropped from 249 street-view images collected from Google Street View. Each word image has a 50-word lexicon.

- **Street View Text Perspective** [30] contains 639 cropped testing images which are specially picked from the side-view angles in Google Street View. Most of them suffer from large perspective distortions. Each image is associated with a 50-word lexicon.

In Section 3.2 and Section 3.3, "50", "1K" and "Full" denote each scene text recognised with a 50-word lexicon, a 1,000-word lexicon and a full lexicon respectively. "None" represents unconstrained scene text recognition without any lexicon.

3.1 Implementation Details

Since our STAR-Net is an extremely deep network, it is difficult to simultaneously optimise the feature extractor and the spatial transformer from scratch. We adopt a pre-training strategy to successfully train the whole STAR-Net. First of all, we train the feature extractor together with CTC on a toy dataset, in which all the text images are tightly bounded and horizontal. This toy dataset is generated using the tools provided by Jaderberg *et al*. [11]. Training on the toy dataset guarantees our feature extractor and CTC understand what kind of text regions is ideal for text recognition. In our experiments, we find this strategy provides good initial parameters for the residue feature extractor and shortens the whole training procedure. We then add a spatial transformer with an affine transformation and train the whole network on the synthetic dataset released by Jaderberg *et al*. [18]. This dataset contains text images with different kinds of deformations. We initialise the parameters of the spatial transformer to represent an identity transformation. After the network is converged, we treat the parameters of this network as our pre-trained parameters. In order to handle more complex distortions, we replace the spatial transformer with one using a more flexible 10-point thin plate spline (TPS) transformation [6]. We use the pre-trained parameters to initialise the network with a TPS transformation, and fine-tune the whole network with a relatively small learning rate. In this paper, all the networks are trained using grey-scale images. The input sizes of our spatial transformer and the following feature extractor are 150×48 and 100×32, respectively. Batch Normalization [17] is used after every convolutional layer. The parameters of our networks are all optimised by Adadelta [40]. Our implementation is based on the publicly available code of Torch7 [8] and Warp-ctc [4].

Architectures Besides STAR-Net, we also evaluate three other network architectures to demonstrate the effectiveness of each component in STAR-Net. These architectures are listed in Table 1 and summarised as follows:

- **CRNN** is proposed by Shi *et al*. [32]. It has seven convolutional layers (arranged as plain convolutional blocks, see Figure 3(b)) and two BLSTM layers.

- **STA-CRNN** introduces the spatial attention mechanism into the architecture of CRNN. The input to the feature extractor in this architecture is the rectified image outputted from the spatial transformer.

Name	SAM	Unit1 (2,2,2,2)	Unit2 (2,2,2,2)	Unit3 (1,2,1,2)	Unit4 (1,2,1,2)	Unit5	BLSTM
CRNN	N	$[3,64]\times 1$	$[3,128]\times 1$	$[3,256]\times 2$	$[3,512]\times 2$	$[3,512]\times 1$	2
STA-CRNN	Y	$[3,64]\times 1$	$[3,128]\times 1$	$[3,256]\times 2$	$[3,512]\times 2$	$[3,512]\times 1$	2
R-Net	N	$[3,64]\times 5$	$[3,128]\times 4$	$[3,256]\times 4$	$[3,512]\times 4$	$[3,512]\times 1$	1
STAR-Net	Y	$[3,64]\times 5$	$[3,128]\times 4$	$[3,256]\times 4$	$[3,512]\times 4$	$[3,512]\times 1$	1

Table 1: Four architectures for scene text recognition. "SAM" represents spatial attention mechanism. The sizes of convolutional filters and channels are shown in brackets with number of layers stacked. After each unit, a max-pooling layer is used except the last one. The width, height, strides of the max-pooling kernel are shown below each unit. Each LSTM has 256 hidden units.

- **R-Net** is a simplified version of STAR-Net, with the spatial attention mechanism removed. It has eighteen convolutional layers and one BLSTM layer in its feature encoder. Compared with CRNN, R-Net is roughly 2.5 times deeper and is optimised with residue learning.

- **STAR-Net** integrates the spatial attention mechanism with residue learning, and have 26 convolutional layers in total. It is by far the deepest model proposed for scene text recognition.

Specifically, we directly use four plain convolutional blocks for the localisation network in our spatial transformer. The filter size, stride and padding size of all the convolutional layers are 3, 1 and 1 respectively. In order to reduce the computational complexity, the channels of these four blocks are 16, 32, 64 and 128, respectively. Each plain convolutional block is followed by a 2×2 max-pooling layer with a stride of 2. A fully connected layer with 256 hidden units is used for outputting all the transformation parameters.

3.2 Results on General Datasets

Experiments are first conducted on four general datasets, namely, ICDAR-2003, ICDAR-2013, IIIT5K and Street View Text, and the results are shown in Table 2. It can be observed that, on both tasks of lexicon-based and lexicon-free recognition, STAR-Net is able to achieve either highly competitive or even state-of-the-art performance. Note that in these datasets, most of the text regions are horizontal but not tightly bounded, and only few of them are severely distorted. The remarkable performance on these four datasets shows the ability of STAR-Net in handling loosely bounded scene texts. From Table 2, we can find that the performance of STAR-Net on both ICDAR datasets is not as good as that on IIIT5K and SVT. One possible explanation is that the proportion of deformed text images in SVT and IIIT5K is higher than that in the ICDAR datasets. This might cause the spatial attention mechanism not being effective on the ICDAR datasets. Besides, there are more blur or low-resolution text images in the ICDAR datasets. This makes our spatial attention mechanism more difficult to locate the text region and rectify it precisely. By comparing the results of CRNN and R-Net, we also observe that significant improvement can be brought by exploiting deep convolutional layers with residue learning. On the other hand, the effectiveness of the spatial attention mechanism is not that obvious in these four general datasets.

Method	IC03			IC13	IIIT5K			SVT	
	50	Full	None	None	50	1K	None	50	None
ABBYY [37]	56.0	55.0	-	-	24.3	-	-	35.0	-
Wang *et al.* [30]	76.0	62.0	-	-	-	-	-	57.0	-
Mishra *et al.* [27]	81.8	67.8	-	-	64.1	57.5	-	73.2	-
Novikova *et al.* [25]	82.8	-	-	-	-	-	-	72.9	-
Wang *et al.* [33]	90.0	84.0	-	-	-	-	-	70.0	-
Bissacco *et al.* [4]	-	-	-	87.6	-	-	-	90.4	78.0
Goel *et al.* [9]	89.7	-	-	-	-	-	-	77.3	-
Alsharif and Pineau [1]	93.1	88.6	-	-	-	-	-	74.3	-
Almazán *et al.* [2]	-	-	-	-	91.2	82.1	-	89.2	-
Lee *et al.* [18]	88.0	76.0	-	-	-	-	-	80.0	-
Yao *et al.* [34]	88.5	80.3	-	-	80.2	69.3	-	75.9	-
Rodriguez-Serrano *et al.* [31]	-	-	-	-	76.1	57.4	-	70.0	-
Jaderberg *et al.* [14]	96.2	91.5	-	-	-	-	-	86.1	-
Su *et al.* [32]	92.0	82.0	-	-	-	-	-	83.0	-
Gordo and Albert [10]	-	-	-	-	93.3	86.6	-	91.8	-
*Jaderberg *et al.* [17]	98.7	**98.6**	**93.1**	**90.8**	97.1	92.7	-	95.4	80.7
Jaderberg *et al.* [15]	97.8	97.0	89.6	81.8	95.5	89.6	-	93.2	71.7
Shi *et al.* [23]	98.3	96.4	90.1	88.6	96.5	93.8	81.9	96.1	81.9
Lee *et al.* [20]	97.9	97.0	88.7	90.0	96.8	94.4	78.4	96.3	80.7
CRNN [22]	**98.7**	97.6	89.4	86.7	97.6	94.4	78.2	96.4	80.8
STA-CRNN	97.1	95.6	89.1	87.3	96.6	92.9	80.1	95.5	80.7
R-Net	98.0	96.5	91.0	89.1	97.2	93.4	83.1	**96.7**	83.2
STAR-Net	96.9	95.3	89.9	89.1	**97.7**	**94.5**	**83.3**	95.5	**83.6**

Table 2: Scene text recognition accuracies (%). All the outputs in * [17] are constrained to a 90K dictionary even when recognising without a pre-defined lexicon.

3.3 Results on SVT-Perspective Datasets

Experiments are next carried out on the Street View Text Perspective dataset (SVT-Perspective). In SVT-Perspective, most of the images suffer from severe perspective distortions. Table 3 shows that our STAR-Net outperforms other state-of-the-art methods on both lexicon-based and lexicon-free recognition. By comparing the results of CRNN and STA-CRNN, we can see the improvement brought by the spatial attention mechanism for highly distorted text images. Besides, it is interesting to notice that even without the spatial attention model, R-Net can achieve a slightly better performance than STA-CRNN. This suggests that having deeper convolutional layers with residue learning can also effectively make the network tolerant to spatial distortions of the scene texts. Some specific examples recognised by our networks are shown in Figure 4. In conclusion, the combining effect of the spatial attention mechanism and deep convolutional layers with residue learning makes our STAR-Net outperform other state-of-the-art methods for loosely bounded scene texts with considerable distortions.

Figure 4: Some examples from SVT-Perspective dataset. "blank" represents the output is empty.

Method	50	Full	None
Wang *et al.* [30]	40.5	26.1	-
Mishra *et al.* [27]	45.7	24.7	-
Wang *et al.* [33]	40.2	32.4	-
Phan *et al.* [30]	75.6	67.0	-
Shi *et al.* [23]	91.2	77.4	71.8
CRNN [22]	92.6	72.6	66.8
STA-CRNN	93.0	80.5	69.3
R-Net	93.0	83.6	70.9
STAR-Net	**94.3**	**83.6**	**73.5**

Table 3: Scene text recognition accuracies (%) on SVT-Perspective dataset.

4 Conclusion

In this paper, we present a novel SpaTrial Attention Residue Network (STAR-Net) for recognising scene texts with considerable distortions. The spatial attention mechanism in our STAR-Net targets at removing the distortions of text in a natural image and producing a tightly bounded text image. This allows the subsequent feature extractor to focus on the rectified text region without being sidetracked by the distortions. Our feature extractor is built upon residue convolutional blocks. It has extremely deep convolutional layers and one recurrent layer (BLSTM) [11] being optimised under the non-parameterised supervision of Connectionist Temporal Classification (CTC). Combining the spatial attention mechanism with the residue convolutional blocks, our STAR-Net is the deepest end-to-end trainable neural network for scene text recognition. Experiments on five public benchmark datasets demonstrate that our STAR-Net can achieve a performance comparable to state-of-the-art methods for scene texts with little distortions, and outperform these methods for scene texts with considerable distortions.

References

[1] https://bitbucket.org/jaderberg/text-renderer/src.

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(12):2552–2566, 2014.

[3] Ouais Alsharif and Joelle Pineau. End-to-end text recognition with hybrid hmm maxout models. *CoRR*, abs/1310.1811, 2013.

[4] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv preprint arXiv:1512.02595*, 2015.

[5] 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*, pages 785–792, 2013.

[6] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):567–585, 1989.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

[9] Vikas Goel, Anadi Mishra, Karteek Alahari, and CV Jawahar. Whole is greater than sum of parts: Recognizing scene text words. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 398–402. IEEE, 2013.

[10] Albert Gordo. Supervised mid-level features for word image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2956–2964, 2015.

[11] Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE, 2013.

[12] Alex Graves. *Supervised sequence labelling*. Springer, 2012.

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[15] Pan He, Weilin Huang, Yu Qiao, Chen Change Loy, and Xiaoou Tang. Reading scene text in deep convolutional sequences. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 3501–3508, 2016.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456, 2015.

[18] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *Workshop on Deep Learning, NIPS*, 2014.

[19] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. In *International Conference on Learning Representations*, 2015.

[20] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Deep features for text spotting. In *Computer Vision–ECCV 2014*, pages 512–528. Springer, 2014.

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[22] 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.

[23] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Mikio Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Jordi Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis-Pere de las Heras. Icdar 2013 robust reading competition. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 1484–1493. IEEE, 2013.

[24] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. *arXiv preprint arXiv:1603.03101*, 2016.

[25] Chen-Yu Lee, Anurag Bhardwaj, Wei Di, Vignesh Jagadeesh, and Robinson Piramuthu. Region-based discriminative feature pooling for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4050–4057, 2014.

[26] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *International Journal of Document Analysis and Recognition (IJDAR)*, 7(2-3):105–122, 2005.

[27] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC 2012-23rd British Machine Vision Conference*. BMVA, 2012.

[28] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[29] Tatiana Novikova, Olga Barinova, Pushmeet Kohli, and Victor Lempitsky. Large-lexicon attribute-consistent text recognition in natural images. In *Computer Vision–ECCV 2012*, pages 752–765. Springer, 2012.

[30] Trung Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013.

[31] Jose A Rodriguez-Serrano, Albert Gordo, and Florent Perronnin. Label embedding: A frugal baseline for text recognition. *International Journal of Computer Vision*, 113(3): 193–207, 2015.

[32] 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. *CoRR*, abs/1507.05717, 2015.

[33] Baoguang Shi, Xinggang Wang, Pengyuan Lv, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. *arXiv preprint arXiv:1603.03915*, 2016.

[34] Bolan Su and Shijian Lu. Accurate scene text recognition based on recurrent neural network. In *Computer Vision–ACCV 2014*, pages 35–48. Springer, 2014.

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.

[36] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016.

[37] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1457–1464. IEEE, 2011.

[38] Tao Wang, David J Wu, Andrew Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

[39] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4049, 2014.

[40] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.