Robust Scene Text Recognition Through Adaptive Image Enhancement

Ye Qian
mf1833053@smail.nju.edu.cn
Yuyang Wang
mf1733066@smail.nju.edu.cn
Feng Su
suf@nju.edu.cn

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

Abstract

Scene text in natural images often has a complex and varied appearance and a variety of degradations, which pose a great challenge to the reliable recognition of text. In this paper, we propose a novel scene text recognition method that introduces an effective, end-to-end trainable text image enhancement network prior to an attention-based recognition network, which adaptively improves the text image and enhances the performance of the whole recognition model. Specifically, the enhancement network combines a novel hierarchical residual enhancement network, which generates and refines pixel-wise enhancement details that are added to the input text image, and a spatial rectification network regularizing the shape of the text. Through end-to-end training with the recognition network in a weak supervision way with word annotations only, the enhancement network effectively learns to transform the text image to a more favorable form for subsequent recognition. The state-of-the-art results on several standard benchmarks demonstrate the effectiveness of our enhancement-based scene text recognition method.

1 Introduction

The rich semantic information conveyed by scene text is of great value in various content-based image applications such as scene understanding, image analysis, video surveillance, and autonomous driving. As a critical step of acquiring text information from natural images, scene text recognition focuses on inferring the underlying character sequence contained in an usually rectangular text region localized in one image by certain text detector, which, on the other hand, is a challenging task due to largely varied appearances, various degradations, and the complex image context of scene text.

In the past years, many scene text recognition methods have been proposed, which can be generally grouped into two categories – character oriented and word oriented. Character oriented methods [23, 30, 32] first localize individual character candidates by connected component analysis or sliding window schemes, and then recognize characters with certain classifiers (e.g., SVM and neural network) and handcrafted or learned features, and finally group characters into words using heuristic rules, clustering algorithms [24], or probabilistic models [30]. However, as character segmentation and recognition are often error-prone, it is usually difficult for these character oriented, bottom-up text recognition methods to achieve optimal performance.

Comparatively, word oriented methods [14, 17, 28] omit the prior segmentation of individual characters and recognize the word as a whole using usually some sequence recognition models capturing linguistic or language knowledge. For example, many of state-of-the-art scene text recognition methods adopted an encoder-decoder framework, in which, the encoder such as a convolutional neural network (CNN) encodes the text image into a sequence of feature vectors, and then the decoder such as a recurrent neural network (RNN) predicts a corresponding sequence of potential character label distributions and further transcribes it into the final character sequence using connectionist temporal classification (CTC) [7, 28] or some attention mechanisms [6, 17]. Through end-to-end training of the whole recognition model, these methods attained significantly improved recognition performance relative to traditional character-oriented ones.

More recently, increasing researches [18, 21, 27, 29, 35, 37] have shifted their focus to the recognition of scene text with irregular shapes such as curved or perspectively distorted, and different techniques like image rectification [21, 27, 29, 37] and 2D attention [18] have been developed to deal with such text. On the other hand, few work has been reported to address another important factor that also often leads to recognition failures of scene text – the degradations of scene text such as low contrast to the background, which are not rare in natural images.

In this paper, we propose a novel scene text recognition method that introduces an effective enhancement network to adaptively improve the input text image before feeding it to an attention-based recognition network, which enhances the performance of the whole recognition model. Figure 1 illustrates the architecture of the proposed text recognition model, which is composed of two end-to-end trainable building blocks: 1) an enhancement network combining spatial rectification and pixel-wise enhancement of the image, and 2) an attention-based recognition network. The key contributions of our scene text recognition method are summarized as follows:

- Besides distortions of text shape, our method considers the degradations of scene text like low contrast as equally important factors causing text recognition failures. Accordingly, our method introduces adaptive text image enhancement into the text recognition model, which helps enhance text cues while suppressing interfering image background and hence effectively improves text recognition accuracy. As far as we know, our proposed enhancement-based recognition scheme has been rarely explored in existing researches on scene text recognition.

- We propose an effective text image enhancement model, which, in addition to spatially rectifying the input text image with a Spatial Transformer Network [13], adaptively enhances the image for subsequent recognition by a novel hierarchical residual enhancement network with attentive feature filtering and fusion based on global context information.

- The proposed enhancement-based text recognition model can be end-to-end trained requiring no additional supervision information than word annotations, and achieves state-of-the-art recognition performance on standard scene text benchmark datasets.

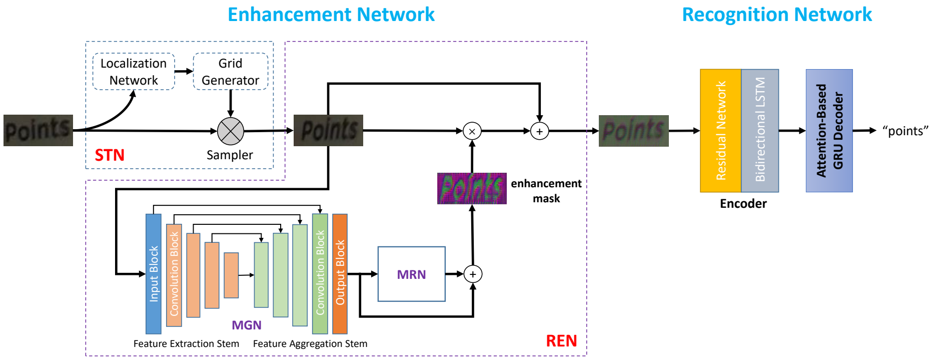

Figure 1: The architecture of the proposed text recognition model.

2 Adaptive Text Image Enhancement

Different from most previous scene text recognition methods, we propose to introduce an adaptive text image enhancement network prior to the recognition network to improve the accuracy of the whole recognition model. As shown in Fig. 1, the proposed text image enhancement model is composed of two main components: a spatial rectification network - STN, and a residual enhancement network - REN.

2.1 Spatial Rectification Network

As one specific form of enhancement to text with irregular shapes such as inclined, curved, or perspectively distorted, we employ a Spatial Transformer Network (STN) [13] as the first module of the enhancement model to spatially rectify the input text image into a relatively regular linear layout.

The STN comprises three components: a localization network, a grid generator, and a sampler. The localization network is usually a convolutional neural network ending with several fully connected layers to predict the x, y coordinates of a set of control points on the input image, which are to be mapped to predefined regular control points on the rectified image. Based on the predicted control points, the grid generator calculates the parameters of a thin plate spline (TPS) [4] transformation and uses it to generate a sampling grid on the input image. Finally, the sampler produces the rectified image by sampling on the grid points with bilinear interpolation.

2.2 Residual Enhancement Network

In this work, we consider the degradation of scene text such as low contrast between text and background as an equally important factor causing a large proportion of failed text recognitions as the distortion of text shape. Accordingly, we propose a residual enhancement network (REN) which adaptively improves the input text image by enhancing the text cues and meanwhile reducing the interference of the background.

As shown in Fig. 1, REN employs a hierarchical residual structure, in which, the predicted enhancement is considered as a residual supplement to the original input image and therefore, once computed, is added to the input image to produce an enhanced representa-

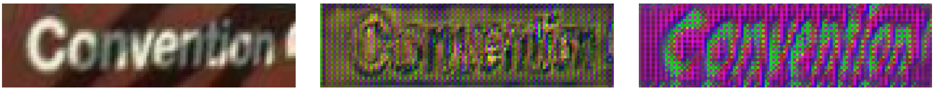

Figure 2: Illustration of an enhancement mask obtained by MGN only (middle) and a refined enhancement mask obtained by the combination of MGN and MRN (right) for an input image (left).

tion of it. Specifically, we employ a two-level framework comprising two cascaded networks *MGN* and *MRN* to predict an *enhancement mask*, which is then element-wise multiplied by the input image to obtain the residual enhancement map to be added to the input image. The first network MGN generates an initial enhancement mask and has a U-shape structure as illustrated in Fig. 1. The next network MRN, which has the same network structure as MGN, further predicts a refinement to be added to the initial enhancement mask which improves the mask with more accurate enhancement details.

Figure 2 shows one example of an enhancement mask predicted by the MGN network only and a refined enhancement mask attained by the combination of the MGN and MRN networks. It can be seen that the refined enhancement mask focuses more accurately on the text region and better suppresses the background.

The U-shape network employed in this work for computing the enhancement mask is composed of two parts – the feature extraction stem and the feature aggregation stem, which gradually aggregate feature maps generated at multiple abstraction levels to accommodate widely varied sizes of characters, as large characters are better depicted by coarser features at higher abstraction levels, while finer features at lower abstraction levels are suitable for depicting small characters.

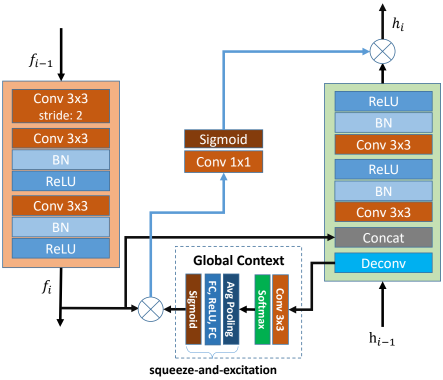

Figure 3: Illustration of a feature extraction block (left) and the corresponding feature aggregation block (right) in the residual enhancement network.

2.2.1 Feature Extraction Stem

The feature extraction stem consists of an input block and four feature extraction blocks. The input block is composed of two convolutional blocks, each consisting of a convolution, a batch normalization (BN), and a rectified linear unit (ReLU). Each feature extraction block, as shown in Fig. 3, first applies a convolution with stride 2 on input feature maps to downsample them to half of the original size, and then employs two convolutional blocks for feature extraction. By cascading four feature extraction blocks, the stem acquires a series of feature representations of the input image, each corresponding to an increasingly higher abstraction level and having a halved resolution compared to the previous representation.

2.2.2 Feature Aggregation Stem

The feature aggregation stem is composed of four feature aggregation blocks and an output block. As shown in Fig. 3, each feature aggregation block first doubles the sizes of the feature maps h_{i-1} generated by the previous feature aggregation block via a deconvolution operation with a stride of 2, and then concatenates them with the feature maps f_i extracted by the corresponding block of the feature extraction stem. Two convolutional blocks are then employed to aggregate features from two stems so as to merge the finer location information encoded in f_i with the semantic information encoded in h_{i-1}.

Finally, the output block consists of a convolutional layer and a sigmoid function and projects the last aggregated feature maps into a three-channel enhancement mask whose values are constrained within $[0,1]$.

2.2.3 Feature Filtering Based on Global Context

In a fully convolutional network, the local receptive field of one convolutional kernel usually cannot capture sufficient global cues for robustly discriminating between text and non-text regions. Therefore, we propose to introduce a global context based attention mechanism to the aforementioned hierarchical U-shape network, which exploits the global information extracted from highly abstracted features to filter and refine the output features of the aggregation stem.

Specifically, as shown in Fig. 3, we first compute a *global context vector* by applying sequentially a convolution, a softmax, and a squeeze-excitation operation [10] on the deconvolved feature maps from h_{i-1} which encode accurate semantic information. The global context vector is then multiplied element-wise to the feature maps f_i of the lower abstraction level, which encode more detailed location information, yielding a set of maps depicting the similarity between the global context vector and f_i. These similarity maps, which are taken as attention cues to help the model to focus on text regions, then undergo a 1×1 convolution followed by a sigmoid function to generate a set of binary weighting maps, which are then used to filter the output of the feature aggregation stem.

2.2.4 Effects of Adaptive Image Enhancement on Text Recognition

Different from a separately-tuned pre-processing module based on some traditional image enhancement techniques which usually enhance both text and non-text regions in the image indiscriminately, the proposed enhancement network is end-to-end trained with the text recognition network, in which the enhancement network is adaptively optimized by propagating the loss at the final text recognition result back. In this way, the enhancement network automatically learns the most favorable way to adapt input images for the text recognition task and effectively increases the accuracy of the recognition results. Meanwhile, the enhancement network requires no extra label information in the training, so that the whole recognition model can be learned in a weak supervision way with only word annotations.

Figure 4 presents some examples of the rectified and enhanced text images by the proposed enhancement model and corresponding recognition results. It can be seen that, given an input text image, the enhancement model produces an image that is usually easier to recognize due to the suppressed smoother background and the enhanced details of the text as well as the spatially rectified text layout, which lead to a more accurate recognition result. Moreover, the proposed enhancement network is sufficiently powerful to deal with text of

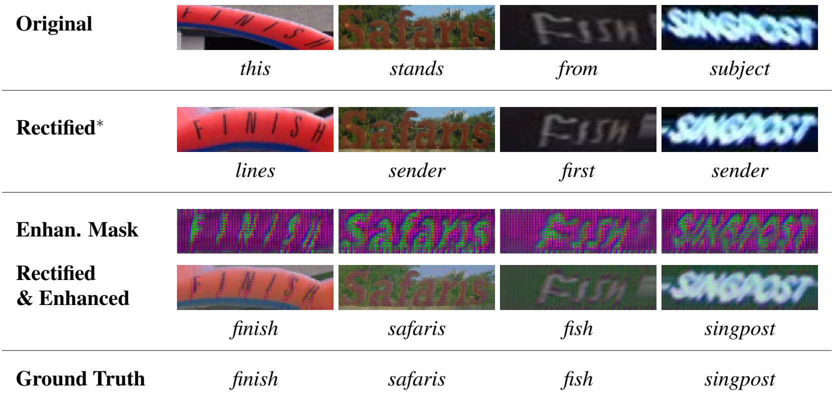

Figure 4: Illustration of the effects of adaptive image enhancement on text recognition results. Row 'Original' presents the original text images. Row 'Rectified*' presents the rectified images by an enhancement model comprising the STN only. Row 'Enhan. Mask' and row 'Rectified & Enhanced' present the predicted enhancement masks and the rectified and enhanced text images by the proposed enhancement model, respectively. Text under the images are the corresponding recognition results. The color representations of the enhanced text image and the enhancement mask are generated by taking the three output maps of the enhancement network (with values normalized to the range $[0,1]$) as the R, G, and B color channels of the image and the mask.

varied sizes owing to its U-shape architecture which captures multi-scale features and the global context information.

3 Attention-Based Text Recognition

On the basis of the proposed text image enhancement model, we employ a relatively standard attention-based sequence-to-sequence recognition model to recognize the text in the enhanced image. The recognition network is composed of an encoder and a decoder as shown in Fig. 1, which is similar to the recognition modules employed in [21, 29].

The encoder first employs a 45-layer residual network [9] consisting of an input block and five residual blocks to extract features from an enhanced text image and finally yield a feature map with a height of 1, which is then sliced along the width axis into a feature sequence. Next, the encoder employs two layers of bidirectional long short-term memory (BiLSTM) to capture long-range dependencies between feature vectors in both forward and backward directions of the feature sequence. The hidden state vectors of the two directions of a BiLSTM are then concatenated as its output feature sequence.

The decoder employs a gated recurrent unit (GRU) with the BahdanauAttention mechanism [1] to decode the encoded sequential features into an output sequence of character label distribution vectors, based on which the decoder further predicts the symbol at each time step and stops processing when it predicts an end-of-sequence token "EOS" [31]. Furthermore, the beam search algorithm is employed in the inference stage, which maintains top-k candidates according to the accumulative scores and finally outputs a relatively optimal character sequence.

3.1 Loss Function

As the image enhancement network is end-to-end trained requiring no additional supervision information, the loss of the whole text recognition model is the same as those employed in most text recognition methods, which is formulated as:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{N}\sum_{j=1}^{NC}\mathbb{I}(\hat{\boldsymbol{y}}_i^j = 1)log(\boldsymbol{y}_i^j) \tag{1}$$

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

4 Experiments

4.1 Dataset

We evaluate our scene text recognition method on following benchmark datasets:

IIIT5K-Words (IIIT5K) [22] contains 3000 cropped text images from Internet, most of which are focused text with regular layout.

ICDAR 2003 (IC03) [20] consists of 251 scene text images, in which the ground-truth text is labeled with its bounding box. We employ the same scheme as used in [32] to discard images consisting of non-alphanumeric characters or containing less than three characters, yielding a dataset composed of 867 cropped text images.

ICDAR 2013 (IC13) [15] inherits most of its samples from IC03 and provides 1015 text images after filtering as applied on IC03.

Street View Text (SVT) [32] contains 647 cropped word images collected from Google Street View for testing, many of which are severely corrupted by noise and blur and may have very low resolutions.

SVT-Perspective (SVT-P) [25] consists of 639 cropped images picked from side-view angle snapshots in Google Street View, which often contain severe perspective distortions.

ICDAR 2015 (IC15) [16] contains 2077 cropped images, with more than 200 text samples being irregular ones such as arbitrarily-oriented, curved, and perspectively distorted.

CUTE80 (CT80) [26] contains 288 cropped natural images and focuses on evaluating recognition performance on curved text.

On all datasets, we employ *word recognition accuracy* as the performance metric, which is defined as $\frac{|C|}{|T|}$ with C and T being the sets of correctly recognized words and all ground-truth words respectively. No lexicon information is exploited in all experiments.

4.2 Implementation Details

The spatial rectification network employs the same configurations as used in [29] and outputs a rectified text image of size 32×112, which is also the input size of the residual enhancement network and the recognition network.

In the residual enhancement network, all convolutional layers use a kernel size of 3×3, except that a 1×1 kernel is used in global-context-based feature filtering as shown in Fig. 3. All deconvolutional layers use a kernel size of 4×4. The number of the output filters

Table 1: Comparison of text recognition performance with several variants of the enhancement network.

Method	Regular Text				Irregular Text		
	IIIT5K	SVT	IC03	IC13	SVT-P	CT80	IC15
Baseline	93.0	87.5	93.9	92.3	78.3	78.5	72.5
STN	93.5	89.5	95.2	93.2	82.2	79.5	75.7
REN$^-$	93.5	88.6	94.8	92.9	82.0	80.2	76.0
STN + REN$^-$	**93.9**	89.8	95.4	93.8	81.2	82.6	76.8
STN + REN$^-$ + GC	93.4	90.1	94.0	94.4	81.2	**84.0**	**77.9**
STN + REN (**Proposed**)	**93.9**	**90.4**	**96.0**	**95.1**	**83.6**	83.7	77.7

employed by the convolutional layers in each block in the feature extraction stem are 32, 64, 128, 256, and 256 respectively, while they are 128, 64, 32, 32, and 3 respectively for the blocks in the feature aggregation stem.

The recognition network employs similar configurations to those used in [21, 29]. In the residual network of the encoder, the input block comprises a 3×3 convolution layer with 1×1 stride, and each of the five residual blocks comprises several residual units, with each unit consisting of a 1×1 convolution and a 3×3 convolution operations. The first two residual blocks each further downsample the feature maps with a 2×2-stride convolution, while the last three residual blocks each employ a 2×1-stride convolution to reduce the sizes of feature maps only in the height direction. The two LSTMs of a BiLSTM layer in the encoder each have 256 hidden units for capturing dependencies in one direction of the feature sequence. The attentive decoder has 512 hidden units and 512 attention units, and recognizes 37 symbols, including 26 lower letters, 10 digits, and 1 EOS token.

We train the proposed enhancement-based text recognition model end-to-end on the 8-million synthetic data by Jaderberg *et al.* [11] and the 6-million synthetic data by Gupta *el al.* [8], which are randomly sampled to form minibatches of size 256. All image samples are resized to 32×112 before being input into the recognition model. The Adadelta [36] optimizer is employed in the training, with the learning rate being initially set to 1.0 and automatically adjusted by the optimizer during the training. The training of the whole recognition model on one NVIDIA Tesla V100 GPU for 10 epochs takes about 40 hours, and the model achieves about 30 FPS in the inference.

4.3 Effectiveness of Adaptive Image Enhancement for Text Recognition

To validate the effectiveness of the proposed adaptive text image enhancement model for scene text recognition, in Table 1, we compare the text recognition performance of combining several variants of the enhancement network with the same recognition backbone. Specifically, the model 'Baseline' skips the prior image rectification and enhancement networks at all, feeding the input text image directly to the recognition backbone. The model 'STN' employs solely the spatial rectification network for enhancing the input image. The model 'REN$^-$' omits the global context block and enhancement mask refinement (i.e., MRN) from the proposed residual enhancement network (REN), enhancing the input image based on the initial enhancement mask without rectification. The model 'STN + REN$^-$' combines the STN and REN$^-$ networks, and the model 'STN + REN$^-$ + GC' further integrates the global context (GC) based feature filtering mechanism into the enhancement network.

Compared to the baseline model, the introduction of either the proposed residual image enhancement module or the spatial rectification module effectively increases the accuracy of the recognition model. Combining both modules helps further enhance the recognition performance on all but one datasets. Meanwhile, integrating the global context block into the enhancement network increases the final recognition accuracy on the majority of datasets, revealing the effect of global context information on feature filtering for text image enhancement. By further introducing enhancement mask refinement, REN attains more accurate residual enhancement representations, which help improve the recognition accuracy on all regular and one irregular text datasets, along with close results on the last two datasets.

The overall enhanced recognition accuracy of the proposed enhancement-based text recognition model on all datasets relative to the baseline demonstrates the effectiveness of adaptive image enhancement for the scene text recognition task. Through end-to-end training with the recognition module, the enhancement network successfully learns to transform the input image in a way leading to better recognition results.

On the other hand, to verify the superiority of the proposed adaptive image enhancement model, which can be jointly learned with the recognition network, over a separate pre-processing scheme (i.e., not end-to-end optimized with the recognizer) employing some traditional image enhancement technique, we replace the proposed REN module in our text recognition model with two common image enhancement techniques, histogram equalization and homomorphic filtering, which act as a pre-processing module and are combined with the retrained spatial rectification and recognition networks. The experimental results show that our enhancement model yields 1.5% to 3.8% higher recognition accuracy on the benchmark datasets compared to the separate pre-processing schemes whose output image may be nonoptimal or even harder for the text recognizer as the non-text parts of the image like the background may also have been enhanced in the absence of the guidance of the recognizer.

4.4 Comparison with State-of-the-Art Text Recognition Methods

We compare our text recognition method with state-of-the-art methods on both regular and irregular scene text datasets in Table 2.

Our method achieves the top recognition accuracy on five of all seven benchmark datasets and the second best accuracy on the other two datasets, among the methods that make use of only word-level annotations in the training as ours. Moreover, compared to the methods that additionally exploited character-level annotations besides word-level ones, including [34] which employed a more complicated spatial rectification model than STN, our method exploiting solely word-level annotations still achieves the highest scores on four of the seven benchmarks and the second best results on the other ones.

The state-of-the-art performance of our method on various benchmarks shows its effectiveness in recognizing scene text with varied shapes and degradations. Given the relatively standard recognition backbone employed in our method, the results of the experiment demonstrate the significant effect of the proposed text image enhancement mechanism on improving the accuracy of the text recognition model. Particularly, compared to those rectification-based text recognition methods [21, 27, 29, 37], our method with a simple STN-based rectification module and the similar recognition backbone achieves overall enhanced recognition accuracy, revealing that, in addition to rectification measures, adaptive image enhancement is also an important and effective measure to improve the recognition of challenging scene text.

Table 2: Recognition accuracy on regular and irregular scene text datasets without utilizing lexicon. The approaches marked with * are trained with both word-level and character-level annotations. In each column, the best performing result is shown in bold font, and the second best result is shown with underline.

Method	Regular Text				Irregular Text		
	IIIT5k	SVT	IC03	IC13	SVT-P	CT80	IC15
Bissacco *et al.* [3]	-	78.0	-	87.6	-	-	-
Jaderberg *et al.* [12]	-	71.7	89.6	81.8	-	-	-
Jaderberg *et al.* [14]	-	80.7	93.1	90.8	-	-	-
Shi *et al.* [28]	78.2	80.8	93.1	86.7	66.8	54.9	-
Shi *et al.* [27]	81.9	81.9	90.1	88.6	71.8	59.2	-
Lee *et al.* [17]	78.4	80.7	88.7	90.0	-	-	-
Cheng *et al.* [6]	87.0	82.8	91.5	-	73.0	76.8	68.2
Luo *et al.* [21]	91.2	88.3	95.0	92.4	76.1	77.4	68.8
Shi *et al.* [29]	93.4	89.5	94.5	91.8	78.5	79.5	76.1
Li *et al.* [18]	91.5	84.5	-	91.0	76.4	83.3	69.2
Zhan *et al.* [37]	93.3	90.2	-	91.3	79.6	83.3	76.9
Wang *et al.* [33]	**94.3**	89.2	95.0	93.9	80.0	**84.4**	74.5
Ours	93.9	**90.4**	**96.0**	**95.1**	**83.6**	83.7	**77.7**
Yang *et al.* * [35]	-	-	-	-	75.8	69.3	-
Cheng *et al.* * [5]	87.4	85.9	94.2	93.3	-	-	70.6
Liu *et al.* * [19]	92.0	85.5	92.0	91.1	78.9	-	74.2
Bai *et al.* * [2]	88.3	87.5	94.6	94.4	-	-	73.9
Yang *et al.* * [34]	**94.4**	88.9	95.0	93.9	80.8	**87.5**	**78.7**
Ours	93.9	**90.4**	**96.0**	**95.1**	**83.6**	83.7	77.7

5 Conclusion

We present a novel enhancement-based scene text recognition method, which adaptively enhances the text image with a spatial rectification network and a hierarchical residual enhancement network before feeding it to a recognition network. Through end-to-end training with the recognition network, the enhancement model automatically learns the most favorable way to transform the image that leads to an enhanced text recognition accuracy. Experiments on various benchmarks demonstrate the effectiveness of our text recognition method for both regular and irregular text.

Acknowledgements. The research was supported by the Natural Science Foundation of Jiangsu Province of China under Grant No. BK20171345 and the National Natural Science Foundation of China under Grant Nos. 61003113, 61321491, and 61672273.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *2018 IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1508–1516. IEEE Computer Society, 2018.

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

[4] Fred L Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6): 567–585, 1989.

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

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

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification. In *Proceedings of the 23rd International Conference on Machine Learning - ICML 2006*, 2006.

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

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[11] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition, 2014.

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

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

[14] 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.

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

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

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

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

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

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

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

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

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

[24] Lukas Neumann and Jiri Matas. Real-time lexicon-free scene text localization and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9): 1872–1885, Sep. 2016.

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

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

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

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

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

[30] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, and Zhong Zhang. Scene text recognition using part-based tree-structured character detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, June 2013.

[31] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

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

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

[34] Mingkun Yang, Yushuo Guan, Minghui Liao, Xin He, Kaigui Bian, Song Bai, Cong Yao, and Xiang Bai. Symmetry-constrained rectification network for scene text recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9146–9155. IEEE, 2019.

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

[36] Matthew D Zeiler. Adadelta: An adaptive learning rate method. *arXiv: Learning*, 2012.

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