A Decoupled Language Model Based on Contrastive Attention Mechanism for Scene Text Recognition

Junwei Zhou
zhoujunwei@iie.ac.cn
Institute of Information Engineering, Chinese Academy of Sciences; School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China

Xi Wang
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
wangxi1@iie.ac.cn

Jiao Dai
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
daijiao@iie.ac.cn

Jizhong Han
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
hanjizhong@iie.ac.cn

ABSTRACT

In recent years, approaches for scene text recognition based on the attention mechanism have achieved amazing success. However, the majority of attention mechanism approaches are coupled, and the majority of ways adhere to the concept of locating the most pertinent image regions. In this paper, we present a language model with a contrastive attention mechanism that is detached from the standard encoder-decoder architecture. First, preliminary text recognition results are obtained based on the encoder-decoder framework; second, we perform the two steps of text prediction in the language model and the calculation of the attention weight of the text to the image, and we not only find the most relevant image area, but also look for the least relevant image area; and finally, the loss function is used to make the model pay less attention to irrelevant areas and more attention to relevant areas. On seven datasets, we evaluated the performance of our model and found that it performed exceptionally well, particularly on the IC13, SVT, and SVTP datasets.

CCS CONCEPTS

• **Applied computing** → **Document capture**.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Junwei Zhou, Xi Wang, Jiao Dai, and Jizhong Han. 2023. A Decoupled Language Model Based on Contrastive Attention Mechanism for Scene Text Recognition. In *2023 9th International Conference on Computing and Artificial Intelligence (ICCAI 2023), March 17–20, 2023, Tianjin, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3594315.3594356

ICCAI 2023, March 17–20, 2023, Tianjin, China

ACM ISBN 978-1-4503-9902-9/23/03.
https://doi.org/10.1145/3594315.3594356

1 INTRODUCTION

Text on images of natural settings includes crucial information such as car plate numbers and store names. To retrieve this important information, scene text recognition was developed as a task.

The scene text recognition task consists primarily of two phases: (1) text detection [1, 28, 41, 42]: locate the location of the text in the image; (2) text recognition [6, 7, 35, 38]: based on the image region discovered during text detection, the text sequence of the region is identified. This article assumes that text detection has been completed and focuses solely on text recognition.

Figure 1: Example of some scene text images

There are two categories for scene text recognition, which are as follows: 1) identification of regular text The regular text is frequently aligned on a single horizontal line and has a transparent backdrop and regular fonts. This type of text is referred to as "regular." 2) Recognizing irregular text Irregular text typically contains a significant deal of information that is hard to differentiate, such as irregular text distribution, various fonts, and fuzzy image quality, among other things. Recognizing irregular text can be challenging because of all of these factors.

Due to the rapid development of deep learning models, the scene text recognition approach based on deep learning has substantially improved its performance in recent years [35, 36]. Images

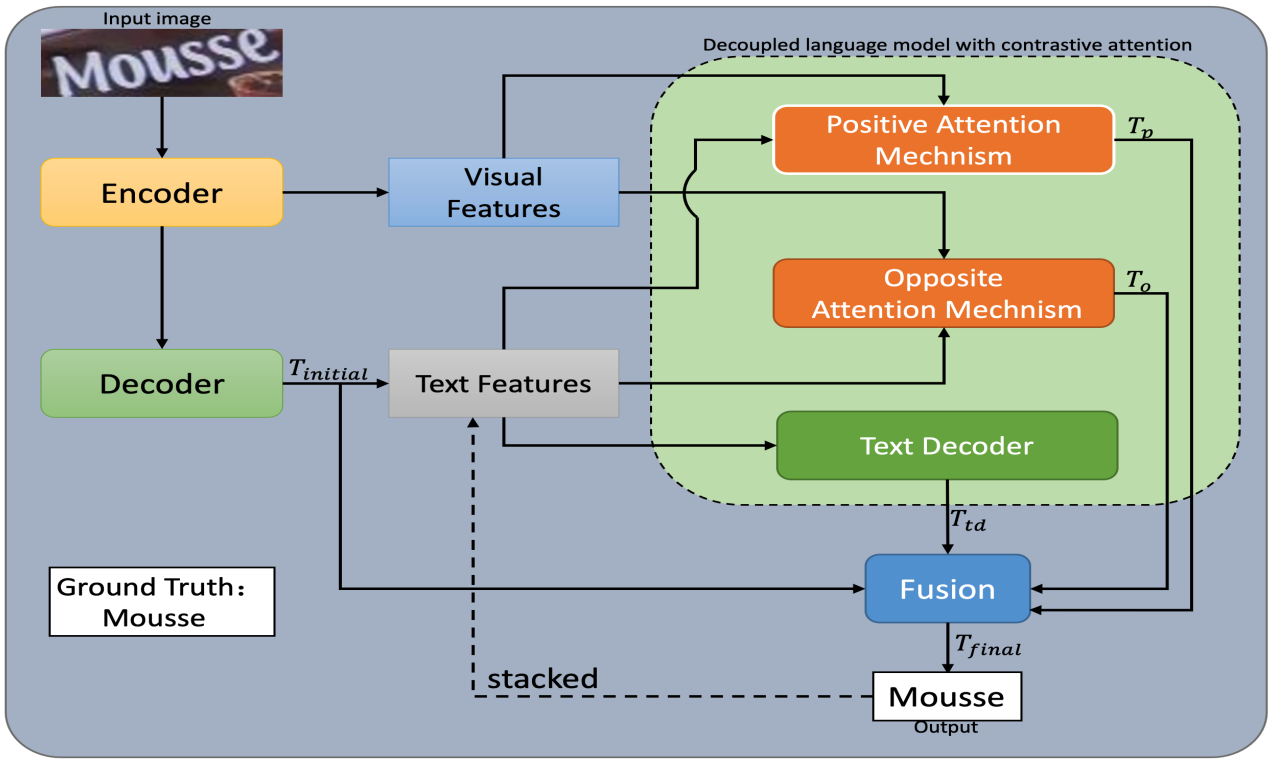

Figure 2: The overall framework of our method.

contain two-dimensional spatial semantics, whereas texts are one-dimensional sequence information. Therefore, the information between images and texts cannot be one-to-one. The key to the present text recognition method is modeling. And can be seen in Figure 1, irregular text has a range of shapes, orientations, colors, sizes, and text appearances. These characteristics may be observed in the text itself. Therefore, the process of obtaining a meaningful text representation for a task involving the recognition of irregular text is still challenging. When attempting to correctly recognize text in scene photographs, a cluttered background as well as a reduction in image quality might lead to significant incorrect predictions.

Significant progress has been made in scene text recognition models equipped with an attention mechanism in recent years [6, 7, 10, 38, 47]. By computing an attention weight, the attention mechanism extracts the image region that is currently the most relevant. However, occasionally, for instance, when the imaging quality of the image is quite low, when the text layout is highly irregular, when numerous portions of the image do not contain text information, etc., these circumstances lead to inaccurate calculations of results.

Some solutions overcome these concerns by utilizing the semanti context knowledge. In the article [51], visual features are first aligned with semantic information, and then a global semantic information inference module is utilized to combine visual and semantic information to produce the result. In [9], the gradient flow between the visual model and the language model is blocked to produce the impact of independent autonomy of the visual model and the language model, and the designed language model achieves more accurate inference results via an iterative mechanism.

In this study, we present a decoupled language model based on a mechanism of contrastive attention. The purpose of our proposed model is to give more attention to relevant visual information and less attention to irrelevant visual information; at the same time, the contrastive attention model and semantic reasoning model can be used independently as functional units and studied separately due to their decoupling. Contributions of this paper mainly include:

(1) We propose a contrastive attention mechanism that gives more weight to relevant information and less weight to irrelevant information when calculating the attention weight of semantic information to visual information.
(2) We present a decoupled language that enables semantic information reasoning and attention weight computation to be conducted independently and that can be optimized by stacking the language model.
(3) Our proposed model obtains superior performance on popular benchmarks, and we present a detailed empirical analysis demonstrating how each component of our model improves scene text recognition performance.

2 RELATED WORK

Traditional methods: Traditional natural scene text recognition approaches rely heavily on the framework for recognizing individual characters. This of approach can be broken down into rely heavily on the framework for recognizing individual characters. This type of approach can be broken down into the following steps:(1) segment the feature of scene text images; (2) send the segmented image features to a single-character classifier intended to recognize each segmentation. (3) combining the recognized single characters in accordance with a set of rules to achieve the final result of recognition [29, 44, 45, 50]. However, due to the presence of several low-quality images in scene text images, bottom-up solutions cannot solve the scene text recognition problem entirely.

Methods based on encoder-decoder framework: Due to the rapid development of deep learning models, the scene text recognition approach based on deep learning has substantially improved its performance in recent years. In the era of deep learning, a visual model is used to encode images into feature spaces by scene text recognition models, the language model decodes image feature into text information. Vision model and language model are correspondingly characterized as encoder and decoder. Currently, text recognition approaches based on deep learning are mostly separated into two categories: (1) methods based on CTC (Connection Temporal Classification) [12, 22] and (2) methods based on sequence learning [6, 7, 34, 38] .

However, since the text is not consistently dispersed and the image quality is inconsistent, additional effort is required to produce better outcomes. The article [26]suggests employing a generative adversarial network [11] to eliminate background information while preserving text content. Text recognition can be improved by utilizing high-resolution images as input; high-resolution images [39] provide consistent high-resolution images. The space transformer (STN) concept [16] provides correction-based navigation. The approach for text recognition [20, 21, 37] greatly enhances text recognition performance.

Attention: The attention method was initially applied to machine translation tasks [2], which can automatically identify the portion of the source sentence that is most relevant to the current word to be predicted. In the field of computer vision, numerous attention-based methods, such as image captioning [49], visual question answering [23], scene text recognition [19], etc., have achieved significant success. As a prediction module, the attention mechanism is frequently paired with a recurrent neural network for scene text recognition tasks. Typically, the input of the attention mechanism is the current instant, the learned history knowledge of the character sequence, and the extracted visual elements of the visual model. By computing an attention weight area, the attention mechanism extracts the image that is currently the most relevant.

A wide variety of scene text recognition systems combined with the attention mechanism have produced impressive results, and there are ongoing efforts to improve the ability to represent information from several perspectives of attention mechanism. The article [7] believes that when the ordinary attention mechanism calculates the attention weight, it expands the visual features in order from left to right and from top to bottom. The article [19] designed a two-dimensional attention mechanism to calculate the degree of correlation between each visual vector and the eight vectors immediately surrounding it, thereby capturing the degree of correlation in the spatial structure. The article [3] proposes an edit probability metric, which calculates the edit distance between the target character sequence and the attention probability distribution prediction sequence, so as to solve the problem of attention drift.

Contrastive Leanring: The core concept of contrastive learning is to decrease the distance between the positive sample and the anchor sample, while increasing the distance to the negative sample. After data improvement from the source sample, it is simpler for the network to learn the common characteristics of multiple samples after data enhancement from the source sample feature [5, 32] . Particularly, the CLIP model has achieved very strong performance on a variety of tasks, drawing even more attention to the contrastive learning technique.

3 METHOD

3.1 Overview

As illustrated in Figure 2, our model incorporates a widely utilized encoder-decoder structure. On the basis of this framework, we offer a language model for performance optimization.

First, given a scene text image I, the encoder extracts the visual feature $V = \{v_1, v_2, \ldots, v_N\}$ of I, and the decoder produces the initial text prediction result $T_{initial} = \{ti_1, ti_2, \ldots, ti_l\}$, in which N is the number of visual features cells and l is the length of $T_{initial}$.

Our suggested language model is then fed V and $T_{initial}$. The positive attention mechanism extracts the most relevant visual features for each character of $T_{initial}$, whereas the opposite attention mechanism extracts the least relevant visual features, and the text decoder performs secondary optimization on the text results.

Finally, we combine $T_{initial}$, T_p, T_o and T_{td} to provide the final text prediction result $T_{final} = \{tf_1, tf_2, \ldots, tf_l\}$. In addition, our suggested language model can be stacked several steps, meaning that the $T_{initial}$ can be replaced with T_{final}.

3.2 Visual and Text Feature Extraction

We employ ResNet [14] and Transformer [43] as encoder and decoder respectively. In addition, the text decoder in out proposed language model is also a Transformer.

Let $X \in R^{r \times H \times W}$ represents visual features, where H and W represent the spatial height and width, and r is the number of channels. And $pos \in R^{l \times h}$ represents the positional encodings of character orders, where l is the length of the character sequence and h is the dimension of the feature. Then V and $T_{initial}$ referred in 3.1 are achieved by transform them, where $N = H \times W$. Besides, V and $T_{initial}$ has the same dimension h.

Different from the conventional Transformer, in the calculation of the attention mechanism, the iterm Q is pos. K and V are both visual features in the decoder, while K and V are both text features in the text decoder.

3.3 Constrative Attention

Positive Attention: The objective of the traditional attention mechanism is to identify the visual information that is most relevant to the current text information. The objective of this paper is to identify the visual feature that match to each character in the text

sequence. This strategy is highly compatible with the typical way of thinking of humans.

Inspired by [43], dot-product attention is applied as a positive attention mechanism :

$$\begin{cases} a_i = QK_i^\top \\ \alpha_i = \dfrac{exp(a_i)}{\sum_{j=1}^{N} exp(a_j)} \\ Q_j = \sum_{i=1}^{N} \alpha_i V_i \end{cases} \tag{1}$$

where Q, K, V denotes query vector, key vectors, and value vectors, respectively. $\alpha_p = \{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ is the attention weight distribution of Q to V. Then the current context information c_p as a vector of element-wise weighted sums of V.

In this paper, our positive attention mechanism focuses on the attention weight distribution of text features to visual features, aims to build the connection between text information and visual information by focusing on the appropriate region of the image. Consequently, Q is the text sequence feature, while K and V are the same two vectors, they are visual features.

Opposite Attention: The attention weights obtained by Equation 1 accurately capture the connection between textual and visual elements, then we design a opposite attention mechanism.

We set the attention weight obtained by Equation 1 as α_p, then the opposite attention weights can be calculated as following:

$$\alpha_o = softmax(1 - \alpha_p) \tag{2}$$

where α_o denotes the opposite attention weights. Therefore, the context information calculation method of the opposite attention mechanism is as follows:

$$c_o = \alpha_o V \tag{3}$$

The function of Equation 2 performs a masking operation. When 1 is subtracted from the positive attention weight, the original maximum value becomes the minimum value, and the minimum value becomes the maximum value. In this way, the most relevant parts of the visual features receive maximum attention under the positive attention mechanism, but are ignored in the opposite attention mechanism. At the same time, the remaining less relevant or irrelevant visual features get more attention under the opposite attention mechanism, so as to be used for contrastive training.

In contrast to positive attention, which summarizes currently relevant visual features, opposite attention summarizes currently irrelevant or less relevant context. They constitute a pair of contrastive information and jointly contribute to the final text generation.

3.4 Fusion

After obtaining the context information c_p and c_o calculated by the forward attention mechanism and the reverse attention mechanism, respectively; and getting the new text features t_{td} calculated by the text decoder, the context information and new text features are concatenated to produce the combined text features:

$$T_{combine} = concat(T_p, T_o, T_{td})W_{combine} \tag{4}$$

where $W_{combine} \in R^{3h \times h}$.

In addition, for the goal of information enhancement, the text features produced by decoder are incorporated into the final text features with a gated mechanism:

$$G = \sigma(concat(T_{initial}, T_{combine})W_f) \tag{5}$$

$$T_{final} = G \odot T_{initial} + (1 - G) \odot T_{combine} \tag{6}$$

3.5 Loss Function

Our model is trained end-to-end, and the multi-task cross-entropy is used as objective function. Then given an image I and its groundtruth, the loss can be formulated as follows:

$$LOSS = L_{initial} + \frac{1}{M}(L_p^i + \lambda L_o^i + L_{td}^i + L_{final}^i) \tag{7}$$

where $L_{initial}$, L_p, L_o, L_{td} and L_{final} are the loss from T_{inital}, T_p, T_o, T_{td} and T_{final}, respectively. L_p^i, L_o^i, L_{td}^i and L_{final}^i are the losses at the i-th step.

It should be pointed out that when calculating L_o , we use the softmin function for text prediction, while other parts utilize softmax function. The text prediction calculation steps are as follows:

$$P_o(y_i|(y_1, y_2, \ldots, y_{i-1})) = softmin(W_o T_o) \tag{8}$$

where $W_o \in R^{h \times h}$. The softmin function is calculated as follows:

$$softmin(t_i) = \frac{e^{-t_i}}{\sum_{j=1}^{l} e^{-t_j}} \tag{9}$$

where $t_i = W_o T_{oi}$ and T_{oi} means the i-th elemens of T_o.

The softmin and softmax functions have opposite function. When we attempt to maximize Po, the opposite attention mechanism searches for the visual feature with the lowest weighting. This indicates that Po can be made higher the less the inverse attention mechanism is connected with visual aspects.

4 EXPERIMENTS

4.1 Datasets

We will validate the performance of our model on seven benchmarks, including four regular datasets and three irregular datasets. The seven benchmarks include IIIT5K-Words (IIIT5K) [27], Street View Text(SVT) [44], ICDAR 2003(IC03) [24], ICDAR 2013(IC13) [18], CUTE80 [33], ICDAR 2015(IC15) [17], and SVT-Perspective(SVTP) [31]. Details of these datasets can be found in [51].

4.2 Implementation Details

Network: we set the size of image to be 32×128, where 32 and 128 represents the height and width respectively. We take Resnet-50 as the encoder and a 4 layers transformer as decoder and text decoder. The dimension h is set to be 512.

Traning: MJSynth (MJ) [15] and SynthText(ST) [13] are the datasets that we use to train. Our model is optimized with the Adam optimizer, where learning rate $lr = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$ and $\epsilon = 10^{-8}$. After five epochs during the training procedure, the learning rating degrades by 0.1.

Implementation: Using the Pytorch framework and two NVIDIA Telsa V100 GPUs, our model is implemented. And the batch size is set to be 192.

Table 1: Comparison of different constrative attention settings.

Base	Text Decoder	Positve Attention	Opposite Attention	IIIT5K	SVT	IC03	IC13	IC15	SVTP	CUTE80	AVG
✓				95.5	91.7	95.4	95.8	84.7	84.8	89.2	91.72
✓	✓			95.9	93.5	95.9	97.3	85.5	89.0	89.2	92.74
✓	✓	✓		96.0	93.0	**97.2**	**97.4**	85.8	88.8	89.6	92.96
✓	✓	✓	✓	**96.1**	**94.4**	96.3	97.3	**85.9**	**90.1**	**90.3**	**93.15**

Table 2: Ablation study of stacking steps. In this study, we set λ to be 0.2 in Equation 7.

staking steps	IIIT5K	SVT	IC03	IC13	IC15	SVTP	CUTE80	AVG
1	**96.0**	**93.5**	96.1	97.1	85.0	86.8	**88.5**	92.46
2	96.0	93.5	**96.4**	**97.8**	**85.7**	**88.8**	88.2	**92.86**

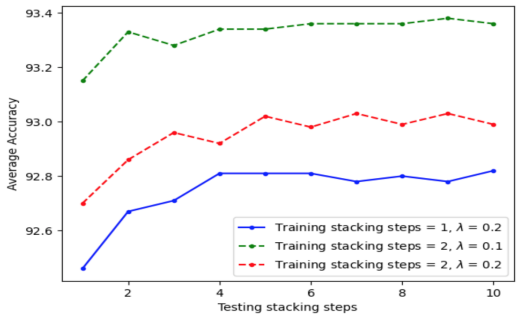

Figure 3: Accuracy of stacking steps in training and testing.

4.3 Ablation Study

To determine the impact of various model components on identification performance, we design a series of ablation experiments. The results of the ablation studies are all evaluated on seven standard benchmarks.

Comparison of constrative attention settings: In this portion of the studies, only the encoder and decoder models will be employed as baseline models. The experimental results of a number of distinct model configurations are presented in the Table 1.

The performance of text recognition is highest when the model contains our proposed decoupled language model based on a contrastive attention mechanism, as shown in Table 1. On both the regular text dataset and the irregular text dataset, the model we developed outperforms the baseline model significantly, with the irregular text dataset showing the most significant performance improvement.

The findings in the table demonstrate that our model can better extract relevant visual elements and increase the distance between related and unrelated visual features.

Comparison of the staked steps of language model: As can be seen from Figure 2, our language model can be stacked in multiple steps to further optimize the recognition results. We present the results of this comparison in Table 2. The results show that when the language model is stacked, the text recognition performance of the model will be improved to a certain extent.

The difference in the number of stacking steps between training and testing was investigated further. The fluctuation of the average precision in Figure 3 implies that: (1) increasing the number of stacking steps in training is beneficial for the performance of the model; (2) applying stacking only in the test can also achieve relatively good results; and (3) the model's performance reaches a relatively saturated state when stacked to more than 4 steps.

Nevertheless, if the number of stacking steps increases during training, the training cost of the model will also grow proportionally; therefore, we set the number of stacking steps at 2 during training to ensure that the model achieves better outcomes while saving training time and money. And from Figure 3, when $\lambda = 0.1$ the model get more better performance.

4.4 Comparisons with state-of-the-art models

In this section, the performance of our method is compared to that of other methods on generic datasets. IIIT5k, SVT, IC03, and IC13 are examples of regular datasets, while CUTE80, IC15, and SVTP are examples of irregular datasets. These findings are presented in Table refresult, with the highlighted value being the second-best performance among all models and the bolded value representing the best performance.

As shown in Table 3, our model achieves the best results for five datasets and the second-best results for two datasets. The effect improvement is especially obvious on the SVT and SVTP datasets, where it is 0.8% and 2.0% higher than the second-best model, respectively. The image quality of these two datasets is blurry, and the encoder is incapable of encoding them effectively. In addition, we discovered that the text decoder in our language model can supplement the initial text features even if the text distribution in the image is irregular and the font is not conventional. Consequently, even without image rectification, our model can achieve the second-best score on the CUTE80 benchmark.

5 CONCLUSION

In this paper, we propose a novel contrastive attention mechanism to make relevant parts of text and imagery to be closer and irrelevant parts farther apart. Simultaneously, the language model we developed is decoupled, allowing it to better learn text information while being unaffected by visual characteristics, and it is also more adaptable. In addition to being applicable to scene text recognition tasks, our method is also applicable to other vision-to-language tasks, such as image captioning, image question answering, etc.

REFERENCES

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF*

Table 3: Compared with other methods.The underlined value represents the second best performance, and the bold value represents the best performance among all models.

Method	Training Data	Regular datasets				Irregular datasets		
		IIIT5K	SVT	IC03	IC13	CUTE80	IC15	SVTP
Luo *et al.* [25]	MJ + ST	91.2	88.3	95.0	92.4	77.4	68.8	76.1
Zhan *et al.* [53]	MJ + ST	93.3	90.2	-	91.3	83.3	76.9	79.6
Li *et al.* [19]	MJ + ST	91.5	84.5	-	91.0	83.3	69.2	76.4
Qiao *et al.* [30]	MJ + ST	93.8	89.6	-	92.8	83.6	80.0	81.4
Wang *et al.* [46]	MJ + ST	94.3	89.2	95.0	95.9	84.4	74.5	80.0
Yue *et al.* [52]	MJ + ST	95.3	88.1	-	94.8	90.3	77.1	79.5
Yu *et al.* [51]	MJ + ST	94.8	91.5	-	95.5	87.8	82.7	85.1
Bhunia *et al.* [4]	MJ + ST	95.2	92.2	-	95.5	89.7	84.0	85.7
Wang *et al.* [48]	MJ + ST	95.8	91.7	-	95.7	88.5	83.7	86.0
Fang *et al.* [9]	MJ + ST	<u>96.2</u>	93.5	-	**97.4**	89.2	86.0	89.3
Tang *et al.* [40]	MJ + ST	**96.3**	<u>93.8</u>	-	96.4	**95.1**	85.4	88.7
Du *et al.* [8]	MJ + ST	**96.3**	91.7	-	97.2	**95.1**	**86.6**	<u>88.4</u>
ours	MJ + ST	**96.3**	**94.6**	**96.6**	**97.4**	<u>91.0</u>	<u>86.2</u>	**90.4**

Conference on Computer Vision and Pattern Recognition. 9365–9374.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. 2018. Edit probability for scene text recognition. In *proceedings of the IEEE conference on computer vision and pattern recognition*. 1508–1516.

[4] Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, and Yi-Zhe Song. 2021. Joint visual semantic reasoning: Multi-stage decoder for text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14940–14949.

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems* 33 (2020), 22243–22255.

[6] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing Attention: Towards Accurate Text Recognition in Natural Images. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2017).

[7] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. 2018. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 5571–5579.

[8] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. 2022. Svtr: Scene text recognition with a single visual model. *IJCAI* (2022).

[9] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7098–7107.

[10] Shancheng Fang, Hongtao Xie, Zheng-Jun Zha, Nannan Sun, Jianlong Tan, and Yongdong Zhang. 2018. Attention and Language Ensemble for Scene Text Recognition with Convolutional Sequence Modeling. *ACM MM* (2018).

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning (ICML)* (2006).

[13] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2315–2324.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

[15] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *IJCV* 116, 1 (2016), 1–20.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. *arXiv* (2015).

[17] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1156–1160.

[18] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. 2013. ICDAR 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1484–1493.

[19] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2019).

[20] Wei Liu, Chaofeng Chen, and Kwanyee K Wong. 2018. Char-Net: A Character-Aware Neural Network for Distorted Scene Text Recognition. *AAAI Conference on Artificial Intelligence (AAAI)* (2018), 7154–7161.

[21] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. 2016. Star-net: a spatial attention residue network for scene text recognition.. In *BMVC*, Vol. 2. 7.

[22] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. 2020. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9809–9818.

[23] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems* 29 (2016).

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

[25] Canjie Luo, Lianwen Jin, and Zenghui Sun. 2019. MORAN: A Multi-Object Rectified Attention Network for scene text recognition. *Pattern Recognition (PR)* 90 (2019), 109–118.

[26] Canjie Luo, Qingxiang Lin, Yuliang Liu, Lianwen Jin, and Chunhua Shen. 2021. Separating content from style using adversarial learning for recognizing text in the wild. *International Journal of Computer Vision* 129, 4 (2021), 960–976.

[27] Anand Mishra, Karteek Alahari, and CV Jawahar. 2012. Scene text recognition using higher order language priors. *BMVC* (2012), 1–11.

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. *Advances in neural information processing systems* 27 (2014).

[29] Lukas Neumann and Jiri Matas. 2012. Real-time scene text localization and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[30] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. 2020. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13528–13537.

[31] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. 2013. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

569–576.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

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

[34] Baoguang Shi, Xiang Bai, and Cong Yao. 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* 39, 11 (2016), 2298–2304.

[35] Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell (TPAMI)* 39, 11 (2017), 2298–2304.

[36] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust Scene Text Recognition with Automatic Rectification. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2016).

[37] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust Scene Text Recognition with Automatic Rectification. *arXiv* (2016).

[38] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2018. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Trans. Pattern Anal. Mach. Intell (TPAMI)* 41, 9 (2018), 2035–2048.

[39] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. 2019. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9086–9095.

[40] Xin Tang, Yongquan Lai, Ying Liu, Yuanyuan Fu, and Rui Fang. 2022. Visual-semantic transformer for scene text recognition. *AAAI* (2022).

[41] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. 2019. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4234–4243.

[42] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiay Jia. 2019. Learning Shape-Aware Embedding for Scene Text Detection. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2019).

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

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

[45] Kai Wang and Serge Belongie. 2010. Word spotting in the wild. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer, 591–604.

[46] Peng Wang, Lu Yang, Hui Li, Yuyan Deng, Chunhua Shen, and Yanning Zhang. 2019. A simple and robust convolutional-attention network for irregular text recognition. *arXiv preprint arXiv:1904.01375* 6, 2 (2019), 1.

[47] Siwei Wang, Yongtao Wang, Xiaoran Qin, Qijie Zhao, and Zhi Tang. 2019. Scene Text Recognition via Gated Cascade Attention. *ICME* (2019), 1018–1023.

[48] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14194–14203.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[50] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. 2014. Strokelets: A learned multi-scale representation for scene text recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (2014), 4042–4049.

[51] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12113–12122.

[52] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*. Springer, 135–151.

[53] Fangneng Zhan and Shijian Lu. 2019. Esir: End-to-end scene text recognition via iterative image rectification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 2059–2068.