Towards Accurate Scene Text Recognition with Semantic Reasoning Networks

Deli Yu[13*] Xuan Li[2*] Chengquan Zhang[2*†] Tao Liu[2]
Junyu Han[2] Jingtuo Liu[2] Errui Ding[2]
School of Artificial Intelligence, University of Chinese Academy of Sciences[1]
Department of Computer Vision Technology(VIS), Baidu Inc.[2]
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences[3]
yudeli2018@ia.ac.cn
{lixuan12, zhangchengquan, liutao32, hanjunyu, liujingtuo, dingerrui}@baidu.com

Abstract

Scene text image contains two levels of contents: visual texture and semantic information. Although the previous scene text recognition methods have made great progress over the past few years, the research on mining semantic information to assist text recognition attracts less attention, only RNN-like structures are explored to implicitly model semantic information. However, we observe that RNN based methods have some obvious shortcomings, such as time-dependent decoding manner and one-way serial transmission of semantic context, which greatly limit the help of semantic information and the computation efficiency. To mitigate these limitations, we propose a novel end-to-end trainable framework named semantic reasoning network (SRN) for accurate scene text recognition, where a global semantic reasoning module (GSRM) is introduced to capture global semantic context through multi-way parallel transmission. The state-of-the-art results on 7 public benchmarks, including regular text, irregular text and non-Latin long text, verify the effectiveness and robustness of the proposed method. In addition, the speed of SRN has significant advantages over the RNN based methods, demonstrating its value in practical use.

1. Introduction

Text has rich semantic information, which has been used in many computer vision based applications such as automatic driving [43], travel translator [38], product retrieval, etc. Scene text recognition is a crucial step in scene text reading system. Although sequence-to-sequence recognition has made several remarkable breakthroughs in the past decades [19, 37, 41], text recognition in the wild is still a big challenge, caused by the significant variations of scene text in color, font, spatial layout and even uncontrollable background.

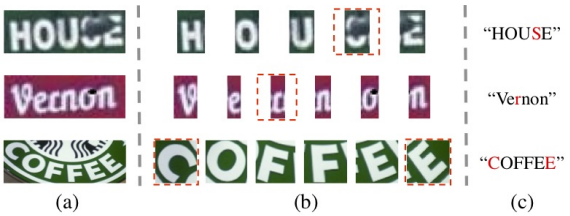
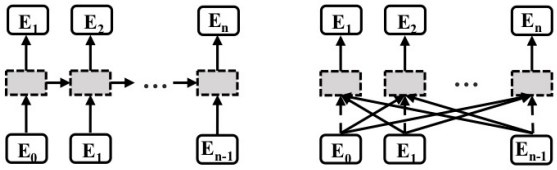


Figure 1. Examples of text in the wild. (a) are some difficult scene text images, (b) are individual characters extracted separately from (a), and (c) are the corresponding semantic word contents. The characters with red dashed boxes in (b) are easy to be misclassified, only based on visual features.

Most of the recent works have attempted to improve the performance of scene text recognition from the perspective of extracting more robust and effective visual features, such as upgrading the backbone networks [5, 21, 31], adding rectification modules [30, 31, 40, 44] and improving attention mechanisms [5, 37, 41]. Nevertheless, it is a fact that, for a human, the recognition of scene text is not only dependent on visual perception information, but also affected by the high-level text semantic context understanding. As some examples shown in Fig. 1, it is very difficult to distinguish each character in those images separately when only visual features are considered, especially the characters highlighted with red dotted boxes. Instead, taking semantic context information into consideration, human is likely to infer the correct result with the total word content.

Unfortunately, for the semantic information, the mainstream text recognition methods consider it in the way of one-way serial transmission, such as [5, 30, 31, 37, 40, 41, 44], which recursively perceive the character semantic information of the last decoding time step, as shown in Fig. 2 (a). There are several obvious drawbacks in this manner: First, it can only perceive very limited semantic context

*Equal contribution. This work is done when Deli Yu is an intern at Baidu Inc.
†Corresponding author.



(a) (b)

Figure 2. Two different manners of semantic context delivery. (a) is one-way serial transmission, and (b) is multi-way parallel transmission.

from each decoding time step, even no useful semantic information for the first decoding time step. Second, it may pass the wrong semantic information down and cause a error accumulation when the wrong decoding is raised at an earlier time step. Meanwhile, the serial mode is hard to be paralleled, thus it is always time-consuming and inefficient.

In this paper, we introduce a sub-network structure named global semantic reasoning module (GSRM) to tackle these disadvantages. The GSRM considers global semantic context in a novel manner of multi-way parallel transmission. As is shown in Fig. 2 (b), the multi-way parallel transmission can simultaneously perceive the semantic information of all characters in a word or text line, which is much more robust and effective. Besides, the wrong semantic content of the individual character can only cause quite limited negative impact on other steps.

Furthermore, we propose a novel framework named semantic reasoning network (SRN) for accurate scene text recognition, which integrates not only global semantic reasoning module (GSRM) but also parallel visual attention module (PVAM) and visual-semantic fusion decoder (VSFD). The PVAM is designed to extract visual features of each time step in a parallel attention mechanism, and the VSFD aims to develop an effective decoder with the combination of visual information and semantic information. The effectiveness and robustness of the SRN are confirmed by extensive experiments, which are discussed in Sec. 4.

The major contributions of this paper are threefold. First, we propose a global semantic reasoning module (GSRM) to consider global semantic context information, which is more robust and efficient than one-way serial semantic transmission methods. Second, a novel framework named semantic reasoning network (SRN) for accurate scene text recognition is proposed, which combines both visual context information and semantic context information effectively. Third, SRN can be trained in an end-to-end manner, and achieve the state-of-the-art performance on several benchmarks including regular text, irregular text and non-Latin long text.

2. Related Work

The research of scene text recognition has made significant progress in the past few years. In this section, we will first review some recent text recognition methods, and summarize them into two categories: semantic context-free and semantic context-aware, according to whether semantic information is used. Besides, some popular context modeling structures will also be presented.

Semantic context-free methods regard scene text recognition as a purely visual classification task, without using any semantic information explicitly. CRNN [29] firstly combined CNN and RNN to extract sequential visual features of a given text image, and then directly fed them into a CTC decoder to predict the best character category of each time step, where CTC [8] only maximized the probability of all the paths that can reach the ground truth according to the visual classification of each position. In order to alleviate the back-propagating computation burden of CTC loss, Xie *et al*. [39] proposed the aggregation cross-entropy (ACE) loss to optimize the statistical frequency of each character along the time dimension, enhancing the efficiency greatly. Inspired by the success of visual segmentation, Liao *et al*. [22] used FCN to predict the character categories of each position by pixel-level classification, and gathered characters into text lines by heuristic rules. However, this method requires expensive character-level annotation. Instead of optimizing the decoding accuracy of each step, Jaderberg [13] directly used CNNs to classify 90k kinds of text image, each of which represented a word. In general, these methods ignore to take semantic context into account.

Semantic context-aware methods try to capture semantic information to assist the scene text recognition. Most of those methods follow the one-way semantic transmission manner, for example, Lee *et al*. [19] encoded the input text image horizontally into 1D sequential visual features, and then guided visual features to attend the corresponding region with the help of semantic information of last time step. As we mentioned before, some of the latest works focus on how to mine more effective visual features, especially for irregular text. In order to eliminate the negative effects brought by perspective distortion and distribution curvature, Shi *et al*. [31] added a rectification module before sequence recognition, in which a spatial transform network [15] with multiple even control point pairs was adopted. Zhan *et al*. [44] employed a line fitting transformation with iterative refinement mechanism to rectify the irregular text image. Furthermore, Yang *et al*. [40] proposed a symmetry-constrained rectification network based on the rich local attributes to generate better rectification results. There are some methods alleviating the challenge of irregular text recognition by enhancing spatial visual features. Cheng *et al*. [6] extracted scene text features in four directions and designed a filter gate to control the contribution of features from each direction. Wojna *et al*. [37] introduced additional encoding of spatial coordinates on 2D

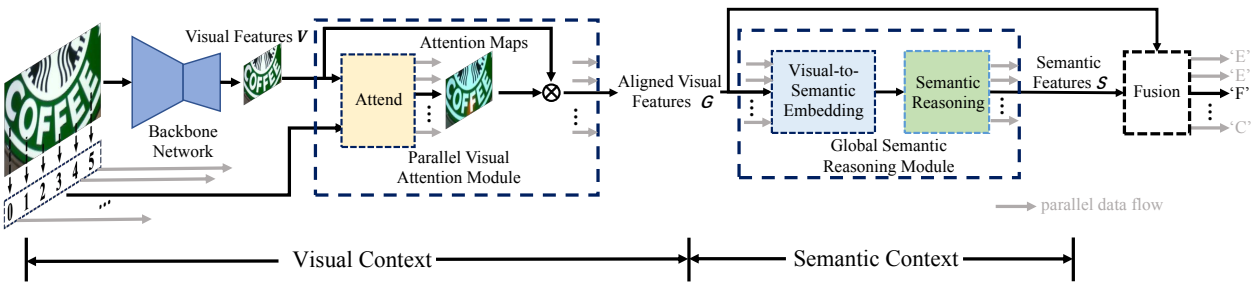


Figure 3. The pipeline of the semantic reasoning network (SRN).

feature maps to increase the sensitivity to sequential order. However, these works do not fully exploit semantic context information, which is exactly what we want to focus on in this paper.

Context modeling structures are designed to capture information in a certain time or spatial range. RNN is good at capturing dependencies of sequence data, but its inherent sequential behavior hinders parallel computation [37] during the training and inference. To solve those issues, ByteNet [16] and ConvS2S [7] directly used CNNs as encoder. These methods can be fully parallelized during training and inference to make better use of the hardware, but cannot flexibly capture the global relations, due to the limitation of receptive field size. Recently, the structure of transformer [24] has been proposed to capture global dependencies and relate two signals at arbitrary positions with constant computation complexity. In addition, transformer has been proved to be effective in many tasks of computer vision [11, 36] and natural language processing [34]. In this paper, we not only adopt transformer to enhance the visual encoding features, but also use the similar structure to reason semantic content.

3. Approach

The SRN is an end-to-end trainable framework that consists of four parts: backbone network, parallel visual attention module (PVAM), global semantic reasoning module (GSRM), and visual-semantic fusion decoder (VSFD). Given an input image, the backbone network is first used to extract 2D features V. Then, the PVAM is used to generate N aligned 1-D features G, where each feature corresponds to a character in the text and captures the aligned visual information. These N 1-D features G are then fed into our GSRM to capture the semantic information S. Finally, the aligned visual features G and the semantic information S are fused by the VSFD to predict N characters. For text string shorter than N, 'EOS' are padded. The detailed structure of SRN is shown in Fig. 3.

3.1. Backbone Network

We use FPN [23] to aggregate hierarchical feature maps from the stage-3, stage-4 and stage-5 of ResNet50 [10] as the backbone network. Thus, the feature map size of ResNet50+FPN is $1/8$ of the input image, and the channel number is 512. Inspired by the idea of non-local mechanisms [4], we also adopt the transformer unit [34] which is composed of a positional encoding, multi-head attention networks and a feed-forward module to effectively capture the global spatial dependencies. 2D feature maps are fed into two stack transformer units, where the number of heads in multi-head attention is 8 and the feed-forward output dimension is 512. After that, the final enhanced 2D visual features are extracted, denoted as V, $v_{ij} \in \mathbf{R}^d$, where $d = 512$.

3.2. Parallel Visual Attention Module

Attention mechanism is widely used in sequence recognition [5, 30]. It can be regarded as a form of feature alignment where relevant information in the input is aligned to the corresponding output. Therefore, attention mechanism is used to generate N features where each feature corresponds to a character in the text. Existing attention based methods are inefficient because of some time-dependent terms. In this work, a new attention method named parallel visual attention (PVA) is introduced to improve the efficiency by breaking down these barriers.

Generally, attention mechanism can be described as follows: Given a key-value set (k_i, v_i) and a query q, the similarities between the query q and all keys k_i are computed. Then the values v_i are aggregated according to the similarities. Specifically, in our work, the key-value set is the input 2D features (v_{ij}, v_{ij}). Following the Bahdanau attention [2], the existing methods use the hidden state H_{t-1} as the query to generate the t-th feature. To make the computation parallel, the reading order is used as the query instead of the time-dependent term H_{t-1}. The first character in the text has reading order 0. The second character has reading order 1, and etc. Our parallel attention mechanism can be summarized as:

$$\begin{cases} e_{t,ij} &= W_e^T tanh(W_o f_o(O_t) + W_v v_{ij}) \\ \alpha_{t,ij} &= \dfrac{\exp(e_{t,ij})}{\sum_{\forall i,j} \exp(e_{t,ij})} \end{cases} \quad (1)$$

where, W_e, W_o, and W_v are trainable weights. O_t is

the character reading order whose value is in the list of $[0, 1, ..., N-1]$, and f_o is the embedding function.

Based on the idea of PVA, we design the parallel visual attention module (PVAM) to align all visual features of all time steps. The aligned visual feature of the t-th time step can be represented as:

$$g_t = \sum_{\forall i,j} \alpha_{t,ij} v_{ij} \tag{2}$$

Because the calculation method is time independent, PVAM outputs aligned visual features (G, $g_t \in \mathbf{R}^d$) of all time steps in parallel.

As some attention maps shown in Fig. 4, the obtained attention maps can pay attention to the visual areas of corresponding characters correctly and the effectiveness of PVAM is verified well.

Figure 4. Attention maps calculated by PVAM.

3.3. Global Semantic Reasoning Module

In this section, we propose the global semantic reasoning module (GSRM) that follows the idea of multi-way parallel transmission to overcome the drawbacks of one-way semantic context delivery. Firstly, we review the probability formula to be maximized in the Bahdanau attention mechanism, a typical RNN-like structure. It can be expressed as the following:

$$p(y_1 y_2 \cdots y_N) = \prod_{t=1}^{N} p(y_t | e_{t-1}, H_{t-1}, g_t) \tag{3}$$

where e_t is regarded as the word embedding of t-th label y_t. At each time step, RNN-like methods can refer to the previous labels or predicted results during the training or inference. So they work in a sequential way since the previous information like e_{t-1} and H_{t-1} can only be captured at time step t, which limits the ability of semantic reasoning and causes low efficiency during inference.

To overcome the aforementioned problems, our key insight is that instead of using the real word embedding e, we use an approximated embedding e' which is time-independent. Several benefits can be made from this improvement. 1) First, the hidden states value of last steps H_{t-1} are able to be removed from the Eq. 3 and thus the serial forward process will be upgraded to a parallel one with high efficiency because all time-dependent terms are eliminated. 2) Second, the global semantic information, including all the former and the latter characters, is allowed

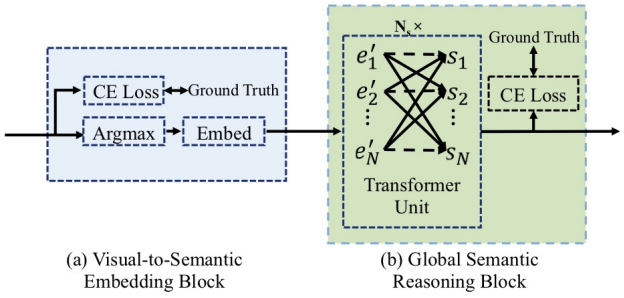


Figure 5. The detailed structure of GSRM.

to be combined together and to reason the appropriate semantic content of the current time. Hence, the probability expression can be upgraded as:

$$\begin{aligned} p(y_1 y_2 \cdots y_N) &= \prod_{t=1}^{N} p(y_t | f_r(e_1 \cdots e_{t-1} e_{t+1} \cdots e_N), g_t) \\ &\approx \prod_{t=1}^{N} p(y_t | f_r(e'_1 \cdots e'_{t-1} e'_{t+1} \cdots e'_N), g_t) \end{aligned} \tag{4}$$

where e'_t is the approximate embedding information of e_t at the t-th time step. The f_r in Eq. 4 means a function that can build the connection between the global semantic context and current semantic information. If we denote the $s_t = f_r(e_1 \cdots e_{t-1} e_{t+1} \cdots e_N)$ and s_t is the t-th feature of semantic information S, the Eq. 4 can be simplified to the following one:

$$p(y_1 y_2 \cdots y_N) \approx \prod_{t=1}^{N} p(y_t | s_t, g_t) \tag{5}$$

Inheriting from the above spirit, we propose the GSRM, by which the function f_r in Eq. 4 is modeled, to make the supposition come true and benefit from it. The structure of GSRM is composed of two key parts: visual-to-semantic embedding block and semantic reasoning block.

Visual-to-semantic embedding block is used for the purpose of generating e', and the detailed structure is shown in Fig. 5 (a). Thanks to the PVAM, the features we get are already aligned to every time step or every target character. The aligned visual features G are fed to a fully-connection layer with softmax activation first, and the embedding loss L_e, where cross entropy loss is utilized, is added to make them more concentrate on the target characters.

$$L_e = -\frac{1}{N} \sum_{t=1}^{N} \log p(y_t | g_t) \tag{6}$$

Next, embedding vector e'_t is calculated based on the most likely output characters of g_t by the $argmax$ operation and an embedding layer.

Semantic reasoning block serves to realize the global semantic reasoning, that is to model the function f_r in Eq. 4. Specially, the structure of GSRM is illustrated in Fig. 5 (b).

Several transformer units are followed with the masked e' allowing the model to perceive the global context information with high efficiency. Meanwhile, both first-order relations and higher-order relations, such as word semantic, can be implicitly modeled by multiple transformer units. Finally, the semantic features of every step is output through this module, which is defined as S, $s_t \in \mathbf{R}^d$, $d = 512$. Meanwhile, the reasoning loss L_r is added on the s. The objective function can be defined as

$$L_r = -\frac{1}{N}\sum_{t=1}^{N}\log p(y_t|s_t) \tag{7}$$

The cross entropy loss is performed to optimize the objective probability from the perspective of semantic information, which also helps to reduce the convergence time. It is noticeable that the global semantic is reasoned in a parallel way in the GSRM, making SRN run much faster than the traditional attention based methods, especially in the cases of long text.

3.4. Visual-Semantic Fusion Decoder

As we discussed in the Sec. 1, it is very important to consider both visual aligned features G and semantic information S for scene text recognition. However, G and S belong to different domains, and their weights for the final sequence recognition in different cases should be different. Inspired by the gated unit [1], we introduce some trainable weights to balance the contribution of features from different domains in our VSFD. The operation can be formulated as the following:

$$\begin{cases} z_t &= \sigma(\mathbf{W_z} \cdot [g_t, s_t]) \\ f_t &= z_t * g_t + (1 - z_t) * s_t \end{cases} \tag{8}$$

where W_z is trainable weight, f_t is the t-th fused feature vector, $t \in [1, N]$. All fused features can be denoted as F, $f_t \in \mathbf{R}^d$, and are used to predict the final characters in a non-recursive manner, the objective function is as follows:

$$L_f = -\frac{1}{N}\sum_{t=1}^{N}\log p(y_t|f_t) \tag{9}$$

Combing all constraint functions in GSRM and VSFD, the final objective function is summarized as follows:

$$Loss = \alpha_e L_e + \alpha_r L_r + \alpha_f L_f \tag{10}$$

where L_e, L_r and L_f represent embedding loss, reasoning loss and final decoder loss, respectively. The weights of α_e, α_r and α_f are set to 1.0, 0.15 and 2.0 to trade off these three constraints.

4. Experiment

4.1. Datasets

There are six Latin scene text benchmarks chosen to evaluate our method.

ICDAR 2013 (IC13) [18] contains 1095 testing images. Using the protocol of [35], we discard images that contain non-alphanumeric characters or less than three characters.

ICDAR 2015 (IC15) [17] is taken with Google Glasses without careful position and focusing. We follow the same protocol of [5] and use only 1811 test images for evaluation without some extremely distorted images.

IIIT 5K-Words (IIIT5k) [26] is collected from the website and comprises of 3000 testing images.

Street View Text (SVT) [35] has 647 testing images cropped form Google Street View. Many images are severely corrupted by noise, blur, and low resolution.

Street View Text-Perspective (SVTP) [27] is also cropped form Google Street View. There are 639 test images in this set and many of them are perspectively distorted.

CUTE80 (CUTE) is proposed in [28] for curved text recognition. 288 testing images are cropped from full images by using annotated words.

4.2. Implementation Details

Model Configurations The details of backbone are introduced in Sec.3.1. For PVAM, the size of W_e, W_o and W_v is 512, and the embedding dim of f_o is 512 in Eq.1. The embedding dim in GSRM is also set to 512. The semantic reasoning block consists of 4 stacked transformer units, where the number of heads is 8 and the number of hidden units is 512. For fair comparison, the same backbone as our SRN is adopted in the CTC, 1D-Attention and 2D-Attention based methods. The number of both attention units and hidden units in 1D-Attention and 2D-Attention are set to 512.

Data Augmentation The size of input images is 64×256. We randomly resize the width of original image to 4 scales (e.g., 64, 128, 192, and 256), and then pad them to 64×256. Besides, some image processing operations [21], such as rotation, perspective distortion, motion blur and Gaussian noise, are randomly added to the training images. The number of class is 37, including 0-9, a-z, and 'EOS'. And the max length of output sequence N is set to 25.

Model Training The proposed model is trained only on two synthetic datasets, namely Synth90K [13, 14] and SynthText [9] without finetuning on other datasets. The ResNet50 pre-trained on ImageNet is employed as our initialized model and the batch size is 256. Training is divided into two stages: warming-up and joint training. At first stage, we train the SRN without the GSRM for about 3 epochs. ADAM optimizer is adopted with the initial learning rate $1e^{-4}$. At joint training stage, we train the whole

pipeline end-to-end with the same optimizer until convergence. All experiments are implemented on a workstation with 8 NVIDIA P40 graphics cards.

4.3. Ablation Study

4.3.1 Discussions about Backbone Network and PVAM

Our SRN utilizes transformer units[34] in the backbone and adds the character reading order information in the PVAM to capture global spatial dependencies. As depicted in Tab. 1, our backbone with transformer units outperforms the one without it on all benchmarks by at least 3% in accuracy, demonstrating that the importance of global visual context captured by transformer unit. As depicted in the in Tab. 1, the gain of using char reading order is obtained in the most public datasets, especially for CUTE. The performance on some easy tasks is slightly improved, since the attention mechanism without this term is actually able to implicitly capture the reading order through data-driven training.

Table 1. Ablation study of backbone and PVAM. "Base" means the backbone; "TU" means the transformer units; "CRO" means the character reading order information.

Method	IC13	IC15	IIIT5K	SVT	SVTP	CUTE
Base	90.0	72.4	87.0	83.6	73.8	80.5
Base+TU	93.0	**77.5**	91.9	87.5	**79.8**	83.6
Base+TU+CRO	**93.2**	**77.5**	**92.3**	**88.1**	79.4	**84.7**

4.3.2 Discussions about GSRM

To evaluate the effectiveness of GSRM in semantic reasoning, we compare the results yielded by the experiments with/without GSRM. Besides, the number of transformer units in GSRM is also explored. As shown in Tab. 2, the GSRM achieves successive gains of 1.5%, 0.2%, 0.8%, 0.3% in IIIT5K and 4.2%, 0.9%, 0.1%, 0.0% in IC15 with the number of Transformer Units set to 1, 2, 4, and 6. This suggests that the semantic information is important to text recognition and GSRM is able to take advantage of these information. Since the performance of 4-GSRM is similar to that of 6-GSRM, the 4-GSRM is adopted in the remaining experiments to preserve controllable computation.

Table 2. Ablation study of GSRM configuration. "n-GSRM" means the GSRM has n transformer units.

	IC13	IC15	IIIT5K	SVT	SVTP	CUTE
no GSRM	93.2	77.5	92.3	88.1	79.4	84.7
1-GSRM	94.7	81.7	93.8	88.5	82.6	**88.9**
2-GSRM	**95.6**	82.6	94.0	91.0	83.9	87.8
4-GSRM	95.5	**82.7**	94.8	**91.5**	**85.1**	87.8
6-GSRM	95.0	**82.7**	**95.1**	90.6	84	86.8

To demonstrate the benefits of global semantic reasoning strategy, we compare our approach with two variants: one only runs forward and the other runs backward to capture one-way semantic information. Moreover, the two typical text recognition methods, CTC and 2D-Attention based methods, are also included in the comparison to prove our superiority to both the existing semantic context-free methods and semantic context-aware methods. As illustrated in Tab. 3, all the semantic context-aware methods outperform the semantic context-free methods (CTC based methods), which highlights the importance of semantic information. Furthermore, the GSRM with global semantic reasoning outperforms those with the one-way semantic reasoning by about 1% in accuracy on most of the benchmarks, verifying the effectiveness of the multi-way semantic reasoning.

Table 3. Ablation study of semantic reasoning strategy. "2D-ATT" means 2D-Attention; "FSRM" and "BSRM" mean forward and backward one-way semantic reasoning module respectively.

	IC13	IC15	IIIT5K	SVT	SVTP	CUTE
CTC	91.7	74.6	91.6	84.5	74.9	81.2
2D-ATT	94.0	77.0	92.7	88.1	78.1	84.3
FSRM	94.7	81.1	**94.9**	89.6	81.7	87.1
BSRM	94.5	81.1	94.3	90.0	82.5	86.8
GSRM	**95.5**	**82.7**	94.8	**91.5**	**85.1**	**87.8**

4.3.3 Discussions about Feature Fusion Strategy

In this paper, we introduce a novel feature fusion strategy, namely gated unit, which is described in Sec. 3.4. In this section, we conduct experiments to compare our VSFD with three different feature fusion methods, including add, concatenate and dot. Tab. 4 indicates that the other three fusion operations will lead to degradation of performance on benchmarks to a certain extent. Thus, the VSFD is utilized in our approach as default.

Table 4. Ablation study of feature fusion strategy.

	IC13	IC15	IIIT5K	SVT	SVTP	CUTE
Add	95.2	81.7	93.8	90.9	84.3	87.8
Concat	95.0	82.0	93.8	**91.5**	82.9	**88.1**
Dot	94.8	81.0	92.0	89.7	84.5	**88.1**
Gated unit	**95.5**	**82.7**	**94.8**	**91.5**	**85.1**	87.8

4.3.4 Analysis of Right/Failure Cases

To illustrate how semantic information helps SRN to improve the performance, we collect some individual cases from the benchmarks to compare the predictions of SRN with/without GSRM. As shown in Fig. 6, for example, because the character "r" is visually similar to character "c" in the image with word "Vernon", the prediction without GSRM wrongly gives the character "c", while the prediction with GSRM correctly infers the character "r" with the help of global semantic context. The character "e" in "sale", the character "r" in "precious", and the character "n" in "herbert" are handled by the same working pattern.

The failure cases of SRN are shown in Fig. 7, which can be divided into two categories: 1) SRN cannot insert or remove characters, but only modifies wrong characters. If the visual features are aligned wrongly or some characters are missed, SRN cannot infer the correct characters, such as

Figure 6. Right cases of SRN with/without GSRM. The predictions are placed along the right side of images. The top string is the prediction of SRN without the GSRM; The bottom string is the prediction of SRN.

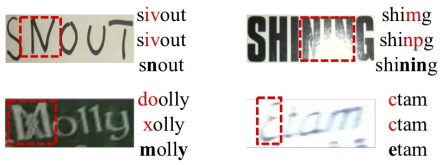


Figure 7. Failure cases of SRN. The predictions and ground-truth labels are placed along the right side of images. The top string is the prediction of SRN without the GSRM; The middle string is the prediction of SRN; The bottom string is the ground-truth.

"snout" and "shining". 2) The SRN may not work when both visual and semantic context fail, since the gain of SRN is mainly due to the complements in visual features and semantic features. When the image suffers from special fonts or low quality and the words in it rarely appear in training data, it is difficult for the SRN to get the visual context and semantic dependencies, as the second line in Fig. 7 shows.

4.4. Comparisons with State-of-the-Arts

The comparison of our method with previous outstanding methods is shown in Tab. 5. We only compare the results without any lexicon, because the lexicon is always unknown before recognition in practical use. The context-aware methods perform better than context-free methods in general, and the proposed SRN achieves superior performance across the six public datasets compared with state-of-the-art approaches with the help of GSRM, which proves that this particular designed module can make better use of semantic information than existing techniques. For regular datasets, we get a 0.2%, 0.4%, 0.9% improvement on IC13, IIIT5K and SVT respectively. The gain of SVT is quite larger than the other two sets, and we claim that semantic information will play a more significant role, especially for recognition of the low-quality images.

Although our method does not take special measures into consideration, such as rectification module, to handle the irregular words, like ASTER [31] and ESIR [44], it is worth noting that the SRN achieves comparable or even better performance on those distorted datasets. As is shown in Tab. 5, there are increase of 4.0% and 2.8% on IC15 and SVTP respectively and comparable results on CUTE, comparing with State-of-the-Arts methods with rectification modules. Similar to the explanation of the gain on SVT, we believe this is mainly due to the fact that global semantic information will be an important supplementation to the visual information in text recognition, and it will show more effectiveness when facing difficult cases.

4.5. Results on non-Latin Long Text

To evaluate the performance on long text, we set up two additional experiments: Attention and CTC based methods with the same configuration. We generate a synthetic long text dataset with the engine in [9], which includes 3 million images. Besides, we also use the training set of RCTW [32] and LSVT [33] as training data. Following the configuration described in Sec. 4.2, we just change the max decoding length N to 50 and the number of classes to 10784. We evaluated our model on ICDAR2015 Text Reading in the Wild Competition dataset (TRW15) [45] by character-level accuracy. TRW15 dataset contains 484 test images. We crop 2997 horizontal text line images as the first test set (TRW-T) and select the images whose length is more than 10 as the second test set (TRW-L).

The results are shown in Tab. 6. Compared with CTC and attention based methods, the proposed approach without GSRM achieved 6.8% and 8.4% boost in TRW-T. Because our method could model 2D spatial information and conquer the error accumulation when the wrong decoding result is raised at a certain time step. Compared with SCCM [42], our SRN achieves 4.9% improvement over the SRN without GSRM, while LM model in SCCM obtains 4.7% improvement. This shows that GSRM could integrate semantic features with visual features very well, which is important to the recognition of long text. Compared with the SRN without GSRM, the improvement of GSRM increases from 4.9% in TRW-T to 6.8% in TRW-L. We can also find that 2D-Attention has a much lower recognition rate in the TRW-L, which is approximately caused by the error accumulation.

As shown in Fig. 8, there are several cases selected from the test set. It's obvious that semantic information can better distinguish two characters, when they are easily confused. For example, the "责" is visually similar to "素", while "素材" is a common Chinese phrase, so the SRN with GSRM correctly infer the character "素".

4.6. Inference Speed

To explore the efficiency of our proposed approach, we evaluate the speed of our method with/without GSRM and compare it with CTC, 1D-Attention and 2D-Attention based recognizers in both the short and long text datasets. The test set is IC15 and TRW-L, of which the average length is 5

Table 5. Comparisons of scene text recognition performance with previous methods on several benchmarks. All results are under NONE lexicon. "90K" and "ST" mean Synth90K and SynthText; "word" and "char" means the word-level or character-level annotations are adopted; "self" means self-designed convolution network or self-made synthetic datasets are used. SRN w/o GSRM means that SRN cuts down GSRM, and thus loses semantic information.

	Method	ConvNet,Data	Annos	IC13	IC15	IIIT5K	SVT	SVTP	CUTE
Semantic context -free	Jaderberg *et al*. [14]	VGG,90K	word	90.8	-	-	80.7	-	-
	Jaderberg *et al*. [12]	VGG,90K	word	81.8	-	-	71.7	-	-
	Shi *et al*. [29] (CTC)	VGG,90K	word	89.6	-	81.2	82.7	-	-
	Lyu *et al*. [25] (Parallel)	ResNet,90K+ST	word	92.7	76.3	94.0	90.1	82.3	86.8
	Xie *et al*. [39] (ACE)	VGG,90K	word	89.7	68.9	82.3	82.6	70.1	82.6
	Liao *et al*. [22] (FCN)	ResNet,ST	word,char	91.5	-	91.9	86.4	-	-
Semantic context -aware	Lee *et al*. [19]	VGG,90K	word	90.0	-	78.4	80.7	-	-
	Cheng *et al*. [5] (FAN)	ResNet,90k+ST	word	93.3	70.6	87.4	85.9	-	-
	Cheng *et al*. [6] (AON)	self,90k+ST	word	-	68.2	87.0	82.8	73.0	76.8
	Bai *et al*. [3]	ResNet,90K+ST	word	94.4	73.9	88.3	87.5	-	-
	Yang *et al*. [41]	VGG,90K+self	word,char	-	-	-	-	75.8	69.3
	Shi *et al*. [31] (ASTER)	ResNet,90K+ST	word	91.8	76.1	93.4	89.5	78.5	79.5
	Zhan *et al*. [44] (ESIR)	ResNet,90K+ST	word	91.3	76.9	93.3	90.2	79.6	83.3
	Yang *et al*. [40] (ScRN)	ResNet,90K+ST	word,char	93.9	78.7	94.4	88.9	80.8	87.5
	Li *et al*. [20] (SAR)	ResNet,90K+ST	word	91.0	69.2	91.5	84.5	76.4	83.3
	Liao *et al*. [21] (SAM)	ResNet,90K+ST	word	95.3	77.3	93.9	90.6	82.2	**87.8**
Ours	SRN w/o GSRM	ResNet,90K+ST	word	93.2	77.5	92.3	88.1	79.4	84.7
	SRN	ResNet,90K+ST	word	**95.5**	**82.7**	**94.8**	**91.5**	**85.1**	**87.8**

Table 6. Recognition accuracies (**character-level**) on non-Latin long text dataset

Method	TRW-T(%)	TRW-L(%)
CASIA-NLPR[45]	72.1	-
SCCM w/o LM[42]	76.5	-
SCCM[42]	81.2	-
2D-Attention	72.2	59.8
CTC	73.8	70.9
SRN w/o GSRM	80.6	77.5
SRN	**85.5**	84.3

深圳篮通光电海限公司　深圳蓝通光电有限公司

合肥创业小吃境训学校　合肥创业小吃培训学校

责材公社　素材公社　配套网络营角　配套网络营销

Figure 8. Right cases for non-Latin long text. There are two predictions under the image. The left string is the prediction of SRN without the GSRM; The right string is the prediction of SRN.

Table 7. Comparion of speed during inference

Method	IC15	TRW-L
CTC	128.6ms	131.8ms
1D-Attention	323.3ms	431.1ms
2D-Attention	338.8ms	486.9ms
SRN w/o GSRM	131.5ms	137.3ms
SRN	191.6ms	216.8ms

and 15 respectively. For a fair comparison, we test all methods with the same backbone network on the same hardware (NVIDIA Tesla K40m). Each method runs 3 times on the test set, and the average time consumed by a single image is listed in Tab. 7.

Benefiting from the parallel framework in SRN, our model with GSRM is 1.7 times and 1.8 times faster than the 1D and 2D-Attention based method in the IC15, and the acceleration will be enlarged to 2.0 times and 2.2 times in the TRW-L. Meanwhile, the computational efficiency of our approach without GSRM is similar to that of CTC-based method, due to its parallelism and simplicity.

5. Conclusion

In this paper, we claim that semantic information is of great importance for a robust and accurate scene text recognizer. Given the characters (Latin or non-Latin) of a text line, we use the GSRM to model its semantic context, which includes both first-order relations among characters and higher-order relations. Integrating with GSRM, we propose a novel end-to-end trainable framework named semantic reasoning network (SRN) for recognizing text in the wild, which also contains backbone network, parallel visual attention module and fusion decoder module. SRN achieves SOTA results in almost 7 public benchmarks including regular text, irregular text and non-Latin long text, and extensive experiments are conducted to show the significant superiority over the existing methods. Additionally, since all modules of SRN are time independent, SRN can run in parallel and is more practical than other semantic modeling methods. In the future, we are interested in improving the efficiency of GSRM, and making it adaptive to CTC-based methods to boost its value in practical applications.

References

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 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. Edit probability for scene text recognition. In *CVPR*, pages 1508–1516, 2018.

[4] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, volume 2, pages 60–65. IEEE, 2005.

[5] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *ICCV*, pages 5076–5084, 2017.

[6] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *CVPR*, pages 5571–5579, 2018.

[7] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, pages 1243–1252. JMLR. org, 2017.

[8] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376. ACM, 2006.

[9] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016.

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

[11] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018.

[12] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*, 2014.

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

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

[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015.

[16] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

[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. Icdar 2015 competition on robust reading. In *ICDAR*, pages 1156–1160. IEEE, 2015.

[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. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE, 2013.

[19] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *CVPR*, pages 2231–2239, 2016.

[20] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *AAAI*, volume 33, pages 8610–8617, 2019.

[21] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[22] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *AAAI*, volume 33, pages 8714–8721, 2019.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[24] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[25] Pengyuan Lyu, Zhicheng Yang, Xinhang Leng, Xiaojun Wu, Ruiyu Li, and Xiaoyong Shen. 2d attentional irregular scene text recognizer. *arXiv preprint arXiv:1906.05708*, 2019.

[26] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

[27] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013.

[28] 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.

[29] 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, 2016.

[30] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *CVPR*, pages 4168–4176, 2016.

[31] 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*, 2018.

[32] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *ICDAR*, volume 1, pages 1429–1434. IEEE, 2017.

[33] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *ICCV*, pages 9086–9095, 2019.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[35] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464. IEEE, 2011.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[37] Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based extraction of structured information from street view imagery. In *ICDAR*, volume 1, pages 844–850. IEEE, 2017.

[38] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *ACM Multimedia*, pages 1500–1508. ACM, 2019.

[39] Zecheng Xie, Yaoxiong Huang, Yuanzhi Zhu, Lianwen Jin, Yuliang Liu, and Lele Xie. Aggregation cross-entropy for sequence recognition. In *CVPR*, pages 6538–6547, 2019.

[40] 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 *ICCV*, pages 9147–9156, 2019.

[41] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. Learning to read irregular text with attention mechanisms. In *IJCAI*, volume 1, page 3, 2017.

[42] Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*, 2017.

[43] Hongyuan Yu, Chengquan Zhang, Xuan Li, Junyu Han, Errui Ding, and Liang Wang. An end-to-end video text detector with online tracking. *arXiv preprint arXiv:1908.07135*, 2019.

[44] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *CVPR*, pages 2059–2068, 2019.

[45] Xinyu Zhou, Shuchang Zhou, Cong Yao, Zhimin Cao, and Qi Yin. Icdar 2015 text reading in the wild competition. *arXiv preprint arXiv:1506.03184*, 2015.