Two-Dimensional Multi-Scale Perceptive Context for Scene Text Recognition

Abstract

Inspired by speech recognition, most of the recent state-of-the-art works convert scene text recognition into sequence prediction. Like most speech recognition problems, context modeling is considered as a critical component in these methods for achieving better performance. However, they usually only consider using a holistic or single-scale local sequence context, in a single dimension. Actually, scene texts or sequence contexts may span arbitrarily across a two-dimensional (2-D) space and in any style, not limited to only horizontal. Moreover, contexts of various scales may synthetically contribute to text recognition, in particular for irregular text recognition. In our method, we consider the context in a 2-D manner, and simultaneously consider context reasoning at various scales, from local to global. Based on this, we propose a new Two-Dimensional Multi-Scale Perceptive Context (TDMSPC) module, which performs multi-scale context learning, along both the horizontal and vertical directions, and then merges them. This can generate shape and layout-dependent feature maps for scene text recognition. This proposed module can be handily inserted into existing sequence-based frameworks to replace their context learning mechanism. Furthermore, a new scene text recognition network, called TDMSPC-Net, is built, by using the TDMSPC module as a building block for the encoder, and adopting an attention-based LSTM as the decoder. Experiments on benchmark datasets show that the TDMSPC module can substantially boost the performance of existing sequence-based scene text recognizers, irrespective of the decoder or backbone network being used. The proposed TDMSPC-Net achieves state-of-the-art accuracy on all the benchmark datasets.

Keywords: Two-Dimensional context, Multi-Scale Perceptive Context, Scene Text Recognition

1. Introduction

Scene text recognition is a critical task for many real-world applications, such as street-sign reading for driverless vehicles, robot navigation, assistive technologies for the blind, etc. Although extensive studies have been carried out in the past few years, text recognition in natural scenes is still challenging, due to several difficulties, e.g., variations of text layout; view distortion, including perspective, curved, and complex geometric deformations; and oriented text.

Inspired by speech recognition, most of the recent algorithms[1][2][3][4][5][6] converted scene text recognition into sequence prediction, which greatly simplifies the problem and leads to promising performance. These sequence-learning-based methods usually employ the encoder-decoder architecture, learning the mapping between the text image and the character sequence. Therein, the encoder converts the input images into a feature sequence, and then captures context information from the sequence. The decoder transcribes the encoded sequence of feature vectors to generate the target strings by using Connectionist Temporal Classification (CTC)[1][2][6] or attention mechanism [4]. Like most speech recognition problems, context dependency is a nontrivial component to provide critical information in these methods, aimed at learning context-aware representation. Recurrent neural networks (RNNs) have been broadly used as context modeling, given their capability in capturing long-range dependency through recurrent computation. However, these RNN context modeling methods are just for 1-D signals, without considering 2-D spatiality of the context. Moreover, these recurrent methods are hard to train, because of gradient dispersion and requiring high computation, due to difficulty in parallel computing.

Recent research has shown that convolutional architectures outperform recurrent networks in the sequence modeling tasks, such as audio synthesis and machine translation [7][8]. Inspired by this, [9][10][5] attempted to adopt stacked

convolutional layers to capture the contextual dependency in scene text images, which overcomes the drawbacks of RNN and serves as a faster alternative to recurrent networks. However, these approaches learn context by using a single fixed receptive field, and regular rectangular convolution kernels, so there is no difference in handling horizontal and vertical context.

Actually, context reasoning in scene text recognition has some intrinsic characteristics, summarized as follows:

(1) Two-dimensional (2-D) context: A character sequence in a text image is essentially extended in the 2-D space, which may appear as an irregular or distorted text. Therefore, encoding context information in the horizontal direction only, or in a single direction, will definitely make the encoder bias towards extracting context features in the dominant direction of the text sequence. However, text patterns may be at any orientation, so they should be decomposed orthogonally and more than one dominant orientation should be considered. In our method, we simultaneously consider the sequence context in both the horizontal and vertical directions, i.e. the 2-D attribute of a context.

(2) Multi-scale context: During the encoding of a text image into a sequence, fixed-size slice segmentation is applied, due to the great difficulty in accurate character segmentation. Each slice is expected to be corresponding to a single character. However, a slice may not be corresponding to a character, because each character may have a different scale and shape. Thus, two basic types of sequential context, namely inter-slice for character and inter-character for word, must be considered for understanding a text string, which involve context dependencies of different sizes. To this end, we propose to capture multi-scale contextual information, with a set of perceptive fields of feasible scales. To the best of our knowledge, almost all the existing encoders are configured with a fixed receptive field, which cannot capture the complete information of a context with varying scales and layouts.

In order to realize the intrinsic characteristics of scene text into a context computing model, we propose a new Two-Dimensional Multi-Scale Perceptive Context (TDMSPC) module for scene text recognition, which can interpret

scene text more effectively and thus, achieve better performance. This module takes any semantic-level feature map as input, and outputs a feature map with the same size, facilitating the merging of 2-D multi-scale perceptive context at any semantic level. As a simple Convolutional Neural Network (CNN) unit, it is compatible with all existing sequence-based recognizers. By inserting this module into a sequence-based recognizer, a unanimous boost in terms of recognition accuracy can be achieved. Moreover, with the TDMSPC module as a core building block, we build up a new scene text recognition network, namely TDMSPC-Net, which achieves state-of-the-art recognition performance on existing benchmark datasets.

2. Related work

Many methods have been proposed for scene text recognition [2][11][10][9][12][13][14][15][16], among which sequence-based methods are especially noteworthy. Examples include [2][1][3][5][15][16], which firstly encoded the input text image into a feature sequence, and then apply the feature sequence to decoders to generate a label sequence. Shi et al. [2] proposed an end-to-end trainable sequence recognition network, which combines CNN and BiLSTM[17] to learn the sequential dependencies, and uses CTC to translate the per-slice prediction into a label sequence. As well, Lee et al. [5] and Cheng et al. [18] constructed an attention-based recurrent network to decode feature sequences and predict labels recurrently. These encoder-decoder frameworks adopt the CNNs to transform images into 1-D sequences, on top of which the recurrent neural network BiLSTM is used to reconfigure these sequence frames for robust sequence transcription. These methods achieve performance gain by integrating contextual information, however, they consider learning context dependencies in one dimension only. In addition, in these existing scene text recognition frameworks, feature extraction and context modeling are processed in two successively separated stages, as summarized in [19], making that the context information is simply integrated into the highest level of semantic feature map. Nevertheless,

both high and low-level features provide rich context information that should be fully utilized. We build the encoding network for the text recognizer by interlacing semantic extraction and context learning from low to high level, thus implementing layer-wise integration of context and semantics.

A text sequence in a scene image may span over a 2-D region, instead of along a single direction. This spatially extended text sequence can be decomposed into two orthogonal directions, and context dependencies can be considered horizontally and vertically. With this insight, the convolution structure is a reasonable option to capture the contextual information, which applies 2-D filters over the entire text image or over the combined maps of all the sequence elements. Recent studies[9][20][21][22][5] show that CNNs are effective to learn the contextual information and demonstrate their advantages over the recurrent connection because of their highly parallelizable convolution computation. Gao et al. [9] applied a sequence-to-map operation to transform a feature sequence into a 2-D map to form the input of stacked multiple CNNs. This can extract the contextual representation of the input sequence to model the global dependencies. Liu et al.[10] developed a multi-scale convolutional encoder to generate a sequence of context vectors and used a scale attention network to select features from the most relevant scales. All these methods capture context from two dimensions unintentionally. However, they treat the context modeling equally along the two orthogonal directions, due to using regular square-shaped kernel filters only. In contrast, our algorithm applies different sets of dilation factor pairs to implement the dilated convolutions, so as to differently handle context modeling in the two dimensions. Furthermore, using dilated convolutions can effectively expand the receptive field, yet the size of the output remains unchanged.

Multi-scale contextual information is always emphasized in modern vision-analysis tasks[23][24][25][26][27], including image classification and semantic segmentation, for performance improvement. However, to the best of our knowledge, no previous works have highlighted multi-scale contextual reasoning to improve scene text recognition. Even though the work of [10] employed multi-

scale information for the scene text encoder, the method simply generated a number of scaled versions of the text image as the input of the network, which introduces expensive computing costs. [21][22] proposed using a multi-scale sliding window, corresponding to different perceptual spans, so as to capture more accurate context information. Except for these two methods, all other existing approaches to scene text recognition capture single-scale context only.

3. TDMSPC Module

We begin by describing the architecture of the proposed TDMSPC module, which is designed to capture 2-D multi-scale perceptive contextual information and to merge them. It is a simple yet effective fully convolutional module, and can be used conveniently to boost the performance of existing scene text recognition methods. The distinguishing characteristics of TDMSPC include: 1) Contextual capturing is realized from the 2-D perspective simply by using dilated convolutions, with different dilation factors to deal with text context in the horizontal and vertical directions. 2) Dilated convolutions are used, with different pairs of dilation factors, to capture contextual information of different scales and shapes. 3) Residual connections are used to merge the 2-D multi-scale context information, without requiring extra parameters. Furthermore, the module can take feature maps of any size as its input, and map them into output feature maps of the same size, which makes it possible to be handily plugged into existing sequence-based text recognizers.

Our module is informed by the recent convolutional architectures for sequential tasks [7] and semantic segmentation [26][23], but is distinct from all of them. Bai et al[7] adopted a one-way causal convolution to look into the past to make a prediction, whereas our model is bidirectional. Furthermore, [7] emphasized covering a global context, whereas our method learns multi-scale context reasoning, and then fuses the multi-scale information. In [26], the context module treated context dependencies in the different dimensions equally. However, our method uses pairs of dilation factors to capture context in the two dimensions

differently, so as to better adapt to the intrinsic characteristics for scene text recognition.

3.1. Architecture

Fig. 1 shows the structure of the proposed TDMSPC, which stacks four dilated convolution layers successively, with different pairs of dilation factors. The dilation factors are increased progressively over the four layers, allowing the module to capture the complete context information in a local-to-global manner. Each of these convolutions is followed by batch normalization[28] and ReLU activation[20], which are not illustrated in the figure for brevity. Finally, by using simple lateral connections, the outputs of all four layers and the input feature maps are added in the element-wise fashion, to produce the output. If the height of the input feature map is less than 4 pixels, the dilation factor for the height or vertical direction is constantly set at 1.

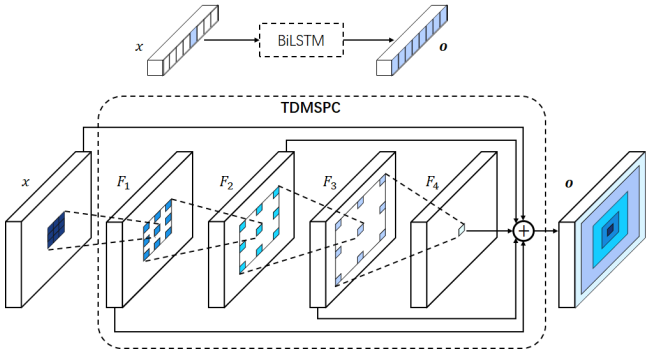

Figure 1: Structure of the TDMSPC module. F_1 is a (1,1)-dilated convolution, and each element of the F_1 output has a receptive field of 3×3. F_2 is a (3,2)-dilated convolution, and each element in the F_2 output has a receptive field of 9×7. F_3 is a (8,4)-dilated convolution, and each element in the F_3 output has a receptive field of 25×15. The number of parameters associated with each layer is identical. Above the TDMPSC also illustrates the traditional BiLSTM method for context learning, which operates on 1-D sequence in a single scale.

3.2. Two-Dimensional Dilation

Dilated convolution can apply the same filter at different regions controlled by different dilation factors, using distinct interval patterns along the horizon-

tal and vertical directions. Therefore, using dilated convolution facilitates the aggregation of contextual information of different scales more efficiently in the 2-D space. The density of aggregation is determined by the dilation factor.

The four layers in the TDMSPC module employ 3×3 convolutions with different dilation factors. Specifically, the dilations along the width dimension are 1, 3, 8 and 23, while those along the height dimension are 1, 2, 4 and 8. Using different dilation factors is due to the fact that a scene text usually has a wider span in the horizontal direction than that in the vertical direction. Actually, these dilation factors serve as hyper-parameters and can be set according to the size of the input feature maps. The dilation rates determined in our method are according to the size distribution of the feature maps of the existing sequence-based text recognizers[4][2][16]. The dilated convolutions operate on the four layers, from the input to the output, using a different pair of dilation factors, i.e. (1,1), (3, 2), (8, 4), and (23, 8), respectively. They capture context dependencies of different scales and shapes in the 2-D text space. Unlike [26], all the convolutions operate in the width and height dimensions, not in the channel dimension. This makes the contextual information conceptually clearer.

Let F_i, where i=1,2,3 and 4, denote the i_{th} layer of the TDMSPC module, and k_i be the convolution filter associated with F_i. For the regular convolution, the output is given as follows:

$$(F_i * k_i)(p^w, p^h) = \sum_{\substack{s^w+t^w=p^w \\ s^h+t^h=p^h}} F_i(s^w, s^h) k_i(t^w, t^h) \tag{1}$$

where (p^w, p^h) represents the coordinates of a point in the output feature map, (s^w, s^h) a point in the input, and (t^w, t^h) a point of the filter.

In our method, we replace the regular convolution with the dilated convolution, with dilation factors (d_i^w, d_i^h) for i = 1, 2, 3 and 4. The dilated convolution $*_{(d_i^w, d_i^h)}$ is defined as

$$(F_i *_{(d_i^w, d_i^h)} k_i)(p^w, p^h) = \sum_{\substack{s^w+d_i^w t^w=p^w \\ s^h+d_i^h t^h=p^h}} F_i(s^w, s^h) k_i(t^w, t^h) \tag{2}$$

The receptive field of the element (p^w, p^h) in the output map of F_i is defined as the set of elements F_i (s^w, s^h) in (2). Denote the size of the receptive field for (p^w, p^h) as $r_i = r_i^w \times r_i^h$. The size of the receptive field r_i of the center unit in the output of F_i can be computed as follows:

$$r_i = (2\sum_{j=1}^{i}(d_j^w) + 1) \times (2\sum_{j=1}^{i}(d_j^h) + 1) \tag{3}$$

for i=1,2,3 and 4.

From (3), the receptive field becomes wider from the first layer to the last layer, which corresponds to using an increasing span for capturing contextual information. The dilation factors are increased monotonically through the layers. However, we do not follow the practice used in dense prediction[26], which increases the dilation factor exponentially with the depth of the network. The reason for this is that dense prediction requires pixel-level accuracy, so the dilation factor is set finer, i.e. 2^i. The dilation factors used in our method are increased more aggressively, according to the feature-map size adopted in existing scene text recognition methods [2][29]. Typically, the original text images are resized to 100×32, which are then reduced to 50×16, 25×8, 25×4, 25×2, and 25×1, by the successive convolutional stages[2][3]. The output feature map of most semantic levels has a width of 25, comparable to the length of text strings. The minimum height can be down to 1, corresponding to 1-D sequences.

The edge units at the output layer will have its range of perceptual context successively decreased due to the boundary effect introduced by convolution. The minimum context range of the edge units in the output of F_i can be formulated as follows:

$$(\sum_{j=1}^{i}(d_j^w) + 1) \times (\sum_{j=1}^{i}(d_j^h) + 1) \tag{4}$$

As above, further analysis of the receptive field of the edge units at the output layer considers multi-focus characteristics of the image text recognition task. In other words, whether central or boundary characters in a text string should be attention-focused. Therefore, all the output units, rather than the center one only, should consider large or global context for reasoning. In our

case, the maximum receptive field of the center element in the output layer is 71×31, while the minimum receptive field size of the element at the border is 36×16. Therefore, the edge output unit of the last layer of the TDMSPC module basically covers the input field globally, given the feature map of the higher semantic stages, with a width of 25 as input.

Assuming that the input length is 25, with the dilation factors set in our method, the center unit of the 2_{nd}, 3_{rd} and 4_{th} layers of TDMSPC covers 12%, 36% and 100%, respectively, of the global context. What we want to emphasize here is that our method can obtain multi-scale contextual information, but the context scale is very discrete. If the scale of the context is to be more fine-grained so as to handle the diversity of irregular text, a feasible way is to implement more combinations of different scales. To this end, we design a recognition network by cascading multiple TDMSPCs continuously or discontinuously, where each module can be plugged into different semantic stages.

4. TDMSPC-Net Recognition Network

Taking the TDMSPC module as a context perception block, we construct a new recognition network, called TDMSPC-Net. We use ResNet-34[30] as our base network, whose architecture is shown on the left of Fig. 2. Inserting a TDMSPC module after each residual block forms the backbone of TDMSPC-Net, as shown on the right of Fig. 2.

The following subsections analyze how the receptive field is affected by stacking the TDMSPC modules and visualize if the feature map adapts to fit irregular text after the module is used. Finally, the GRU attention decoder is briefly described for the sake of completeness.

4.1. More Fine-grained Receptive Field

Superficially, the module takes a feature map as input, and produces a feature map of the same size as output. The input and output have the same form. Inside the TDMSPC module, context of different ranges is captured, and

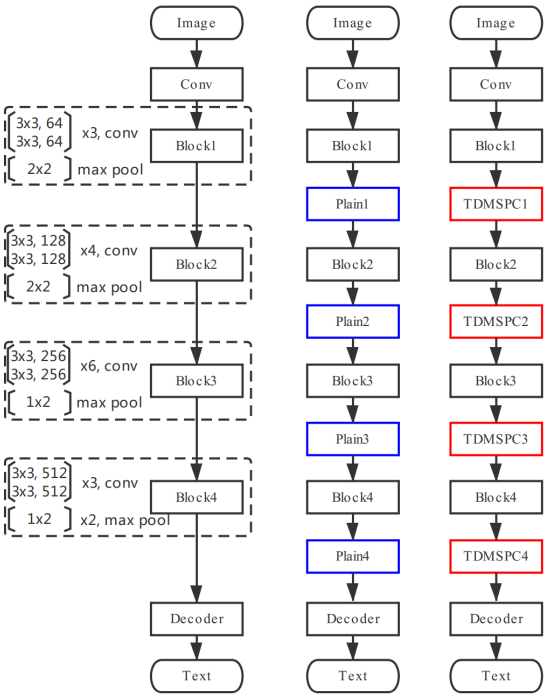

Figure 2: Architecture of the recognition networks. **Left**: The ResNet34 model, denoted as 'Block$_{1,2,3,4}$' in Fig. 7. **Middle**: 'PlainN' defined in Section 5 and used in Tab. 2, and **Right**: the proposed TDMSPC-Net. Inside the dotted box, the structure of each block is given, and more details can be referred to [30].

then merged. Assuming that the input is denoted as x, and the output is o, we formulate the mapping between them as follows:

$$o = x + \sum_{i=1}^{l} (F_i *_{(d_i^w, d_i^h)} k_i)(p^w, p^h), l = 4 \tag{5}$$

where all the symbols follow the definition defined in Section 3. Taking the center unit at the F_i output as an example, we analyze how the receptive fields are combined to produce finer scales.

Using one TDMSPC module obtains a collection of receptive field scales $R = r_0, r_i, i \in 1, ..., l$, where $r_0 = 1 \times 1$ corresponds to the input feature map itself. The output of the module assembles all the scales in R. By stacking multiple TDMSPCs, the receptive fields of different scales are combined to produce more

scales. Consider that two TDMSPC modules are cascaded and x is the input. The output of the first TDMSPC module is denoted as o_1, and that of the second module as o_2. The mappings are then given as follows:

$$o_1 = x + \sum_{i=1}^{l} (F_i^1 *_{(d_i^w, d_i^h)} k_i^1)(p^w, p^h), l = 4 \tag{6}$$

$$o_2 = o_1 + \sum_{i=1}^{l} (F_i^2 *_{(d_i^w, d_i^h)} k_i^2)(p^w, p^h), l = 4 \tag{7}$$

where F_i^1 and F_i^2 denote the i_{th} layer of the first and the second TDMSPC modules, respectively. k_i^1 and k_i^2 are the convolution filters associated with layers F_i^1 and F_i^2, respectively. The first module obtains a collection of receptive field scales $R^1 = \{r_0^{x,1}, r_i^{x,1}, i \in \{1, ..., l\}\}$ for the input x, and the number of scales is denoted as $|R^1|$. In $r_0^{x,1}$ or $r_i^{x,1}$, x represents the input feature map of the module and the '1' implies the first module. Similarly, the second module has a set of receptive field scales $R^2 = \{r_0^{o_1,1}, r_i^{o_1,1}, i \in \{1, ..., l\}\}$, whose input is o_1. The number of scales is denoted as $|R^2|$. It is easy to find that $|R^1|$,$|R^2| \sim O(l)$. However, when two TDMSPC modules are stacked, a more accurate definition of the receptive fields should be given. The receptive field of an element (p^w, p^h) in the output of the j_{th} layer of the second module is the set of elements in input x, instead of o_1, and the size of the receptive field, i.e., $r_{j,i}^{x,2} (i = 0, 1, ..., l)$, of the center unit in the output of F_j^2 is the number of corresponding input elements in x. $r_{j,i}^{x,2}$ is given as follows:

$$r_{j,i}^{x,2} = (2(\sum_{m=1}^{j} d_m^{2,w} + \sum_{n=1}^{i} d_n^{1,w}) + 1) \times (2(\sum_{m=1}^{j} d_m^{2,h} + \sum_{n=1}^{i} d_n^{1,h}) + 1) \tag{8}$$

It is easy to see that the j_{th} layer of the second module outputs $l+1$ different receptive field scales, as $i = 0, 1, ..., l$. In summary, the second module aggregates to provide a collection of receptive field scales, $R = \{r_{j,i}^{x,2}, i, j \in \{0, 1, ..., l\}\}$, where $|R| \sim O(l^2)$. Therefore, the number of receptive-field scales, after stacking the two modules, is exponentially expanded with l, which is much more than that obtained by using a single TDMSPC. Likewise, as in Section 3, a similar example is shown in the following. Assuming that the length of the input is 25,

and aggregating all the receptive-field scales of the center unit of all the layers in the second module, the degree of contextual coverage will be 4%, 12%, 20%, 36%, 44%, 68%, and 100% of the global context. For brevity, we only give the calculation along the width dimension in this example.

A similar derivation can be generalized to obtain the set of receptive-field scales when more modules are stacked. When the TDMSPC modules are inserted into the different convolution stages, the combined effect of the receptive-field scales becomes more complicated, which makes the explicit derivation intractable. This is because the conventional convolution performs striding or pooling, which essentially changes the input semantics to the module. In spite of this, the feasible and diversified effect of combining receptive field scales is still implicit by plugging multiple modules into different convolution stages.

In summary, introducing multiple TDMSPC modules to a recognition network will provide more fine-grained receptive fields. This means that more patterns of the context dependencies will be learned for reasoning.

4.2. Visualizing the Fitted Feature Map

To demonstrate that the feature maps learned by using the TDMSPC module can adapt better to diversified irregular image text, we follow the method of [31] to visualize the output feature maps.

Fig. 3 compares the feature maps of PlainN (described in Section 5.4) and those of TDMSPC-Net at different stages. Two conclusions can be drawn. The first one is that stacking more TDMSPC modules can make the distribution of strong response in the learned feature map gradually fit with the distribution of character regions in the text better. The second one is that the feature map from each stage of TDMSPC-Net fits to the text regions better compared to that from the corresponding stage of PlainN, in particular at the higher semantic stages. Fig. 4 illustrates the output feature maps at Stage 3, i.e. all the feature maps are extracted from Plain3 or the TDMSPC3 output. This figure shows more examples, and a similar conclusion can be reached that our TDMSPC modules can adapt the feature maps well to irregular text, including character layout,

orientation and shape. In contrast, the feature map from the plain convolution is not well adapted to irregular text.

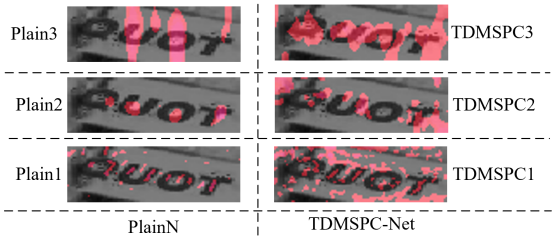

Figure 3: Visualization and comparison of the feature maps from PlainN and TDMSPC-Net at different semantic stages

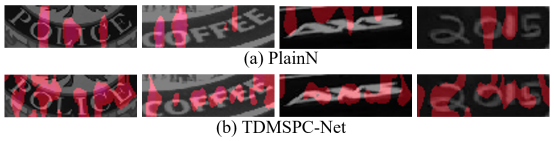

Figure 4: Comparison of the feature maps extracted from the last 2-D semantic layer of (a) PlainN and (b) TDMSPC-Net.

4.3. Attention-based Decoder

After our recognizer is enhanced with layer-by-layer context integration using the TDMSPC module, we obtain an enriched feature map, denoted as $V = \{v_i, i = 1, ..., L\} \in R^{L \times D}$, where D is the channel number and L is the sequence element number. Following most of the existing recognizers[29][15][32], we further adopt attention-based GRU (gated recurrent units) decoders to translate the context enhanced features V into the character sequence $y = (y_1, y_2, ..., y_T)$.

At step t, the decoder starts by computing a vector of attention weights, $\alpha_{t,i}$, as follows:

$$s_{t,i} = S^T(tanh(Wv_i + Uh_{t-1})) \tag{9}$$

$$\alpha_{t,i} = \frac{exps_{t,i}}{\sum_{j=1}^{L} exp(s_{t,j})} \tag{10}$$

where S^T, W, U are the parameter matrices, and h_{t-1} is the last GRU hidden state. Then, we can obtain the weighted sum of the sequential feature vectors, which focuses on the most relevant features:

$$c_t = \sum_{i=1}^{L} \alpha_{t,i} v_i \tag{11}$$

After that, the GRU hidden state is updated and the probability distribution $p(y_t)$ is estimated as follows:

$$h_t = GRU(y_{t-1}, h_{t-1}, c_t) \tag{12}$$

$$p(y_t) = softmax(Y^T h_t) \tag{13}$$

where Y^T is also learnable parameters. Following [29] and [16], we exploit a bidirectional decoder with the 'fractional pickup' training method to further improve our attention GRU decoder.

5. Experiments

This section gives the implementation details and evaluates the performance of the proposed TDMSPC module, as well as the recognition network TDMSPC-Net.

5.1. Datasets

We evaluate the performance using the case-insensitive word accuracy. Two synthetic datasets, Synth90k[33] and ST[34] are used for training. Unless specified, otherwise, all the models in the following experiments are trained using a single Synth90k dataset. Seven benchmark datasets are used for testing, which are grouped into 'Regular' datasets (IIIT5K, SVT, 1C03, IC13) and 'Irregular' datasets (IC15, SVT-P, CUTE). To be noted, 'Regular' and 'Irregular' in Tab. 1, Tab. 2, Tab. 4 and Tab. 5 indicate the average accuracies on these two group of datasets. Following is a brief introduction to the datasets.

IIIT5K[35] was collected from Google image search-es using query words that potentially return text images. It has 3,000 cropped word images for testing.

SVT[36] was collected from Google Street View. The test set consists of 647 word images, some of which are noisy, blurry, or of low-resolution.

IC03[37] contains 867 text images, which are cropped out from 251 scene images and do not contain any nonalphanumeric characters or text that has less than three characters.

IC13[38] is the successor to IC03, so most of its data is inherited from IC03. It contains 1,015 cropped text images, in which words shorter than 3 characters are included.

IC15[39] contains 2,077 text images, most of which are noisy, blurry and rotated, and some are also of low resolution.

SVT-P[40] contains 645 test images, which are specially picked from the side-view angles in Google Street View. Most of them suffer from a large perspective distortion.

CUTE was proposed in [41], focusing on curved text. It contains 80 high-resolution images taken in natural scenes, from which the annotated words are cropped to form a test set of 288 text images.

5.2. Implementation Details

We implemented our network on Pytorch. Experiments were performed on a computer equipped with Intel i7-7700K, 64G RAM and a NVIDIA GTX 1080Ti GPU. The networks are optimized with the stochastic gradient descent (SGD) algorithm, with the initial learning rate 0.1 and momentum 0.9, and the learning rate was decayed by 10% at 0.6 million and 0.8 million iterations, respectively. The size of the training mini-batch is 64. The training is stopped after 0.85 million iterations.

5.3. Ablation Study

In this section, we investigate the effectiveness of the multi-scale receptive field fusion of the TDMSPC module. We also analyze how the plugging of the TDMSPC modules affects the recognition performance.

Setting of dilated factor pairs. We fixedly set the dilation factor along the height dimension as 1, 2, 4 and 8, which is exponentially increasing dilation factor adopted in most existing work[7][23][26]. We adjust dilation factor along the width dimension, combining with dilation factor in the height to form different dilation factor pairs to explore how the setting affect the recognition performance. Here, we only adjust the dilation factor in the width direction because width direction is the principle extending direction for most scene text images. Tab. 1 list different setting of dilation factor pairs, which are denoted as R1, R2, R3 and R4, respectively. Herein, 'R3' is the adopted setting in our paper. 'MaxField' and 'EdgeField' in this table indicate the size of the receptive field of the center unit and edge unit, as computed using equation (3) and (4), respectively. By comparing 'MaxField' or 'EdgeField' and the size of input feature map, we know the scales of context range of TDMSPC module. For clarity and simplicity, we report the average recognition accuracy on regular and irregular datasets, respectively, as well as on all the datasets. They are separately denoted as 'Regular', 'Irregular' and 'Average' in the Tab. 1.

From the table, 'R1' is a traditional dilated convolution, which handles two dimensions equally. It gives the lowest 'Regular', 'Irregular', and 'Average' accuracy of 84.5%, 65.2%, and 77.7%, which are 4.1%, 5.6%, and 4.7% lower than optimal 'R3', respectively. These results imply the remarkable advantage of handling two dimensions differently in TDMSPC compared with handling them equally. Different from our setting 'R3', 'R2' increases the dilation factor not so aggressively, thus can't cover the input field globally, according to the size distribution of the feature maps of the existing sequence-based text recognizers[4][2][16]. It is observed that 'R2' obtain an 'Average' accuracy of 81.5%, which is 0.9 percentage points lower than that of 'R3'. As for 'R4', much larger dilation factor in the width direction makes context range much larger than input field, which results in a significant decrease in the recognition performance.

In summary, it is optimal to designing the dilation factor to make our TDMSPC module to capture the complete context information in a local-to-global

manner. Also, the idea of two-dimensional dilation promotes the significant increase of scene text recognition performance.

	Dilated factor pairs(W, H)	MaxField	EdgeField	Regular	Irregular	Average
R1	(1,1),(2,2),(4,4),(8,8)	31	16	84.5	65.2	77.7
R2	(1,1),(3,2),(5,4),(8,8)	35	18	87.9	69.8	81.5
R3	(1,1),(3,2),(8,4),(23,8)	71	36	**88.6**	**70.8**	**82.4**
R4	(3,1),(8,2),(14,4),(29,8)	109	55	85.4	67.2	80.0

Table 1: Results of different setting of dilation factor pairs. 'H' indicates the dilation factor in the height direction, and 'W' indicates the dilation factor in the width direction. 'Average' indicates the average accuracies on all the seven benchmark datasets (IIIT5K, SVT, 1C03, IC13, IC15, SVT-P, CUTE).

Multi-Scale Receptive Field Fusion. We keep the TDMSPC module that is after Block1, and remove all the other TDMSPC modules in TDMSPC-Net (Fig. 2 Right) to obtain a reduced version. The lateral connections are adjusted to simulate the fusion of different receptive scales inside the TDMSPC, as is illustrated in Fig. 5.

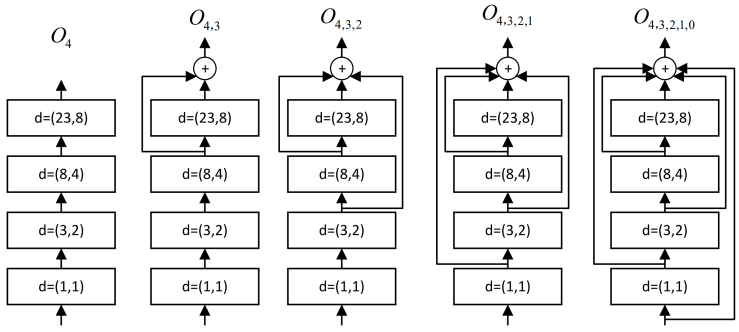

Figure 5: Different schemes of fusing multi-scale receptive fields inside the TDMSPC module. O_n (n=0,1,2,3,4) denotes the output of the n_{th} layer inside TDMSPC. Especially, O_0 is the input of TDMSPC. $O_{4,3,2,1,0}$ represents fusing the output of all the layers along with the module's input. Other symbols, including $O_{4,3,2,1}$, etc., follow similar interpretations.

Fig. 6 compares the performance of five different fusion schemes of the receptive field scales, based on three irregular datasets, SVT-P, IC15 and CUTE.

The results show that the accuracies become higher, when more receptive field scales are merged. Specifically, $O_{4,3,2,1,0}$ outperforms all the other fusion schemes, while using a single scale of receptive field, i.e. O_4, obtains significantly lower accuracy. This proves that the multi-scale context learning and fusion yield a performance gain on the three irregular datasets.

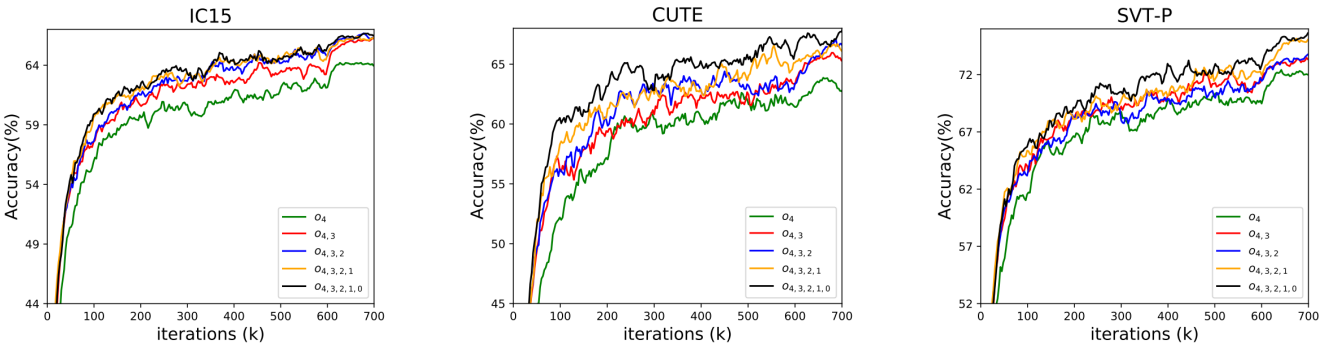

Figure 6: Ablation study of different multi-scale receptive field fusion schemes.

Plugging TDMSPC modules. With the basal ResNet-34 convolution network (Fig. 2 Left), we plug the TDMSPC modules into its different semantic layers. Specifically, we use 'Block_1+' to represent inserting one TDMSPC after Block1, '$\text{Block}_{1,2}$+' means inserting one TDMSPC module after Block1 and Block2, respectively, and so forth. '$\text{Block}_{1,2,3,4}$' means that no TDMSPC is inserted, i.e. a pure ResNet-34 is used as the encoder network. TDMSPC-Net represents the full version with one TDMSPC plugged after each residual block of ResNet-34. In addition, 'Block_4++++' indicates that four TDMSPC modules are continuously inserted after the highest semantic level, i.e. the fourth residual stage of ResNet-34. The designed 'Block_4++++' is to compare the difference between the way of context enhancement coding centrally implemented at the highest semantic level and the way of layer-wise interleaved coding of semantics and context. Two regular datasets, SVT and IIIT5K, and two irregular datasets, IC15 and CUTE, are selected for this part of the experiment. Fig. 7 summarizes all the results. We can see that $\text{Block}_{1,2,3,4}$, without any TDMSPC inserted, performs the worst compared to that with TDMSPC modules insert-

ed. When more and more of the proposed TDMSPC module is inserted into the low-to-high semantic layers of ResNet-34, the recognition performance persistently improves. When the context module is plugged, respectively, after the four residual stages of ResNet-34, our encoder of the TDMSPC-Net recognizer is formed, which achieves the highest accuracy on all the benchmark datasets. Comparing Block_4++++ with TDMSPC-Net, the accuracy of the former is significantly lower than that of the latter, demonstrating that the proposed stage-wise interleaved coding of semantics and context benefits the recognition performance. This idea differs from that adopted in all the existing scene recognizers, in which the context coding is implemented only at the highest layer, e.g. in the Convolutional Recurrent Neural Network (CRNN) methods[4][2][16].

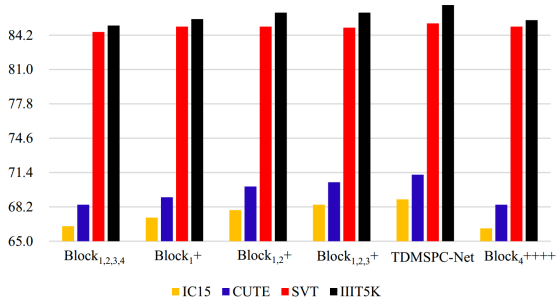

Figure 7: Ablation study of the effect of plugging TDMSPC modules.

5.4. Insight into Performance Gain

As proved previously, the TDMSPC module and its stage-wise insertion into different ResNet-34 semantic levels contribute to the performance improvement of our scene text recognizer. However, it may be questioned if the performance gain is attributed to the number of parameters, the network depth, or the resulting additional computation cost, rather than the TDMSPC module itself and the interleaved semantic-context encoding design. To this end, we design a counterpart, called PlainN, with the same number of convolution layers and model parameters, as TDMSPC-Net. In PlainN (in Fig. 2 Middle), Plain1-4

Method	TDMSPC-Net	PlainN	ResNet-50	ResNet-101
No. of Parameters, layers and Flops				
Para(10^7)	2.9	2.9	2.8	4.7
Layers	49	49	49	100
FLOPS(10^9)	4.2	4.2	3.9	7.7
Accuracy on benchmark datasets				
IIIT5K	**87.1**	83.5	84.5	85.6
SVT	85.3	83.2	84.5	**85.5**
IC03	**93.0**	91.6	91.7	92.5
IC13	**91.3**	88.3	89.3	90.3
IC15	**68.9**	63.9	65.7	68.0
SVT-P	**77.2**	70.8	73.3	74.9
CUTE	**71.2**	66.0	68.1	69.8
Regular	**88.6**	85.6	86.5	87.5
Irregular	**70.8**	65.6	67.6	69.7

Table 2: Recognition accuracies (%) on seven datasets. Models are trained with Synth90k.

are formed by removing all the lateral connections inside a TDMSPC module, and replacing 3×3 dilated convolutions with the regular 3×3 convolutions. Furthermore, we choose two mature deep residual networks, i.e., ResNet-50 and ResNet-101, to serve as the encoder and adopt the same attention GRU decoder to construct the baseline recognizers. Tab. 2 tabulates the results. From the table, TDMSPC-Net has exactly the same number of parameters and convolution layers, as well as computation cost, as PlainN. However, TDMSPC-Net achieves substantially higher accuracies on all the benchmark datasets. We can observe that a performance gain of about 5% and 3% can be achieved on the irregular datasets and regular datasets, respectively. Compared with TDMSPC-Net, ResNet-101 has 62% more parameters, twice as many layers, and 83% more computation. Nevertheless, TDMSPC-Net still achieves significant performance advantages. From the results, TDMSPC-Net achieves an average accuracy of

88.6% on regular datasets, and 70.8% on irregular datasets, both of which are 1.1% higher than that of ResNet-101. In summary, the excellent recognition accuracy of TDMSPC-Net is not attributed to the increase in the number of parameters, network depth or computation cost, but due to the proposed 2-D multi-scale context fusion, implemented in the TDMSPC module, and the interleaved semantic-context encoding structure of the TDMSPC-Net recognizer.

5.5. Boosting the Performance of Existing Recognizers

We selected two state-of-the-art sequence-based scene text recognizers, CRNN[2] and MORAN[29], to verify the boosting effect by plugging the proposed TDMSPC modules. CRNN uses VGG as its backbone network, and adopts CTC as its decoder, without using a spatial transformer to rectify text images. MORAN[42] uses ResNet[30] as its backbone, and adopts attention decoder, with the use of the spatial transformer, MORN, for text rectification. As the structure of these two recognizers are very different, they are good choices for evaluating the performance boosting effect of TDMSPC.

Methods	CRNN		MORAN	
	Origin	TDMSPC+	Origin	TDMSPC+
IIIT5K	83.1	**83.9**	82.5	**83.5**
SVT	80.2	**82.2**	82.1	**82.8**
IC15	62.8	**63.0**	64.1	**65.3**
CUTE	61.8	**63.2**	63.9	**64.9**

Table 3: Boosting effect of TDMSPC modules (%)

Both CRNN and MORAN use BiLSTM to capture 1-D sequence dependencies in their encoders. Whereas in this experiment, we remove all the BiLSTM layers in the encoding part, and insert the TDMSPC modules to capture context information. Specifically, TDMSPC is inserted immediately after those convolution stages that have 64, 128 and 256 output channels, and with the channel number increased at the next stages. All the inserted TDMSPC modules are operated on 2-D feature maps, emphasizing 2-D contextual learning at multiple

semantic levels. Tab. 3 tabulates the results, where 'TDMSPC+' stands for the boosting scheme, while 'Origin' stands for the original methods, without making any changes.

The results show that 'TDMSPC+' unanimously out-performs 'Origin'. This implies that TDMSPC can effectively replace BiLSTM for extracting context information and can achieve performance boosting

5.6. Comparison with State-of-the-Art Methods

We collected all the available published results of deep-learning-based scene text recognition methods for comparison. In view of the fact that some recognizers were trained on the Synth90k dataset only, while others were trained on the combined Synth90k and ST, we carried out the comparative experiments for these two cases. The results are summarized in Tab. 4 and Tab. 5. In addition to TDMSPC-Net, we construct a multi-stage scene text recognition system, called TDMSPC-System, by attaching the spatial transformer MORN, before our recognition network.

As shown in Tab. 4 and 5, TDMSPC-Net and TDMSPC-System outperform all the listed deep-learning-based methods on all the datasets, with the exceptions of CUTE in the two tables and IC13 in Tab. 5. TDMSPC-System performs well, on par with TDMSPC-Net, on the benchmark datasets, which indicates that adding an extra spatial transformer at the front end of TDMSPC-Net is not always necessary for helping to improve the accuracy.

We compare TDMSPC-Net with RARE[3], STAR-Net[4], Char-Net[12], ASTER[16], MORAN[29], ESIR[13], AEG[32], TextSR[43] and ScRN[15], which are specifically designed for recognizing distorted text. These methods employ a spatial transformer, such as STNs[3], MORN, etc., in their encoders. From the recognition results in Tab. 4 and Tab. 5, we found that TDMSPC-Net, without using any rectifier, outperforms all the above nine models by a large margin on almost all the public databases. The exceptional cases, including TDMSPC-Net and ESIR, achieved an accuracy of 71.2% and 72.1% on CUTE, respectively, in Tab. 4; TDMSPC-Net, AEG and ScRN obtained accuracies of 93.3%, 95.3%

Method	RN	IIIT5K	SVT	IC03	IC13	IC15	SVT-P	CUTE	Regular	Irregular
CRNN[2]	No	81.2	82.7	91.9	89.6	-	66.8	-	84.6	-
RARE[3]	Yes	81.9	81.9	90.1	-	-	71.8	-	-	-
R^2AM[5]	No	78.4	80.7	88.7	90.0	-	-	-	82.4	-
STAR-Net[4]	Yes	83.3	83.6	89.9	89.1	-	73.5	-	85.4	-
Char-Net[12]	Yes	83.6	84.4	91.5	-	60.0	73.5	-	-	-
S-SAN[10]	No	85.2	85.5	92.9	90.3	65.7	74.4	-	87.3	-
ACSM[9]	No	81.8	82.7	89.2	88.0	-	-	-	84.2	-
SCCM[22]	No	81.6	76.5	84.5	85.2	-	-	-	82.1	-
SCAN[21]	No	84.2	85.0	92.1	90.4	-	-	-	86.7	-
ASTER-B[16]	Yes	83.2	81.6	92.4	89.7	-	75.4	67.4	85.6	-
ESIR[13]	Yes	82.9	85.9	-	89.1	-	75.8	**72.1**	-	-
TDMSPC-Net	No	**87.1**	85.3	**93.0**	**91.3**	68.9	77.2	71.2	**88.6**	70.8
TDMSPC-System	Yes	86.4	**86.7**	92.7	90	**69.8**	**77.7**	70.1	88.1	**71.5**

Table 4: Recognition accuracies (%) on seven datasets. These models are trained with Synth90k.'RN' means rectification network.

and 93.9%, respectively, on IC13, in Tab. 5. TDMSPC-Net and ScRN achieved an accuracy of 84.7% and 87.5%, respectively, on CUTE, in Tab. 5. It is noteworthy that those methods using a rectification network are particularly prone to achieving prominent results on specific databases. In contrast, TDMSPC-Net focuses on realizing 2-D multi-scale context learning and fusion based on the characteristic of context dependencies, achieving consistently excellent performance on various databases. This is convincingly proven by the observations that the average accuracy of TDMSPC-Net on 'Regular' and 'Irregular' datasets outperforms all the methods equipped with a rectification network, as shown in Tab. 5.

For those datasets, such as IC15, and SVT-P, characterized with many irregular text images, the gain in terms of accuracy with TDMSPC-Net is more prominent. For example, the accuracy of TDMSPC-Net in Tab. 4 is up to

Method	RN	IIIT5K	SVT	IC03	IC13	IC15	SVT-P	CUTE	Regular	Irregular
EP[44]	No	88.3	87.5	94.6	94.4	-	-	-	90.3	-
FAN[18]	No	87.4	85.9	94.2	93.3	-	-	-	89.4	-
AON[45]	No	87.0	82.8	91.5	-	68.2	73.0	76.8	-	70.1
S-SAN[10]	No	91.5	89.6	94.9	93.8	73.4	81.6	-	92.2	-
MORAN[29]	Yes	91.2	88.3	95.0	92.4	68.8	76.1	77.4	91.7	71.2
ASTER[16]	Yes	93.4	89.5	94.5	91.8	-	78.5	79.5	92.8	-
EPAN[46]	No	91.9	88.9	95.0	94.5	73.9	79.4	82.6	92.5	75.9
ESIR[13]	Yes	93.3	90.2	-	91.3	-	79.6	83.3	-	78.1
AEG[32]	Yes	94.6	90.4	95.3	**95.3**	77.4	82.8	81.3	94.3	78.9
TextSR[43]	Yes	92.5	87.2	93.2	91.3	75.6	77.4	78.9	91.8	76.3
ScRN[15]	Yes	94.4	88.9	95.0	93.9	78.7	80.8	**87.5**	93.8	80.0
TDMSPC-Net	No	94.9	90.6	**95.7**	93.3	**80.7**	**85.1**	84.7	94.2	82.0
TDMSPC-System	Yes	**95.2**	**90.7**	95.4	93.7	80.6	85.0	86.5	**94.4**	**82.1**

Table 5: Recognition accuracies (%) on seven datasets. Models are trained with Synth90k and ST. 'RN' means rectification network.

68.9% on IC15, which is 3.2% higher than the second-ranked S-SAN. Similarly, TDMSPC-Net achieves an accuracy of 77.2% on SVT-P, while the accuracy of E-SIR is 75.8%, 1.4% less than that of our TDMSPC-Net. Similar conclusions can be drawn from Tab. 5. Specifically, the most striking example is that TDMSPC-Net achieves a high accuracy of 85.1% on SVT-P, which is 5.7%, 5.5%, 2.3%, 7.7% and 4.3% higher than the recent methods EPAN, ESIR, AEG, TextSR and ScRN, respectively. This analysis further proves that our TDMSPC-Net adapts well to irregular text images, due to the use of a simple 2-D multi-scale context learning and fusion scheme. More experiment results on the Clean Benchmark Datasets and further trade-off analysis of accuracy-speed and accuracy-memory are included in the Appendix, providing a baseline for the future research.

Figure 8: Examples of correct recognition results.

5.7. Exemplar analysis of recognition results

Some correct results based on our method are illustrated in Fig. 8. As can be seen, our method demonstrates an excellent ability for recognizing irregular texts, including curved shape, perspective distortion and arbitrarily-oriented placement, which are common, but challenging in the scene recognition task. Furthermore, we show that the proposed method is capable of recognizing text with large variations in the aspects of illumination, blur, text font, color, etc. These superior results are mainly due to our method's strong ability to lever-age multi-scale context in a 2-D manner.

Figure 9: Examples of false recognition results.'GT' stands for the groundtruth annotation, and 'Pred' denotes the predicted results

Some failure cases are presented in Fig. 9. There are a variety of reasons for failure, including: 1) heavily occluded, 2) overly artistic or distorted, 3) blurry and low-resolution text images, all of which are hardly recognizable to human eyes. This will motivate the future research to solve these cases.

6. Conclusions

In this paper, we have presented a new TDMSPC module, which is a simple yet effective design, by employing two-dimensional multi-scale context perception and fusion. This fully convolutional module can act as a faster alternative to the recurrent network BiLSTM for context modeling in the task of scene text recognition. This module takes any semantic-level feature map as input, and outputs a feature map with the same size as the input, and therefore can be handily plugged into any existing sequence-based scene text recognizers, to boost their performance. Furthermore, we propose to interleave the semantic extraction and context learning, thus implementing gradual integration of context and semantics, from low to high level. Based on this, we build a new, enhanced context encoding network, together with an attention GRU decoder to form our text recognizer TDMSPC-Net. The excellent performance of TDMSPC-Net proves the validity, indicating that both high and low-level feature maps provide rich context information, which should be fully utilized. The experiments on the scene text recognition benchmarks demonstrate that TDMSPC-Net achieves superior or highly competitive performance, compared with the state-of-the-art algorithms. Actually, the advantages of our recognizer are more prominent for the irregular text, demonstrating that the challenges brought by the irregular text can be handled effectively by considering the multi-scale and spatiality of context, without requiring any rectification network to explicitly transform the text into a regular one.

Appendix A. Experimental Results on the Cleansed Bench-marks

We found noisy (incorrect) labels existing in the seven benchmark datasets, which make us confusing on the performance analysis of the proposed method. Thanks to the work in [47], all the cleansed labels are open-sourced, which can be used for the following supplementary evaluations in Tab. 6. It can be observed that the ratio of incorrect label in CUTE dataset is particularly high. The new test accuracy of our proposed TDMSPC-Net on the cleansed CUTE is up to 88.2%, which is higher than that currently reported best result of 87.5%. We are hoping that our reported results here can be used for comparison with other further researches.

	TDMSPC-Net	TDMSPC-Net+Clean
IIIT5K	94.9	95.3
SVT	90.6	90.7
IC03	95.7	95.7
IC13	93.3	93.4
IC15	80.7	81.7
SVT-P	85.1	85.6
CUTE	84.7	88.2

Table 6: The experimental results on the cleansed benchmark datasets.

Appendix B. Analysis of Trade-offs

In order to provide a thorough analysis on the proposed TDMSPC-Net in terms of accuracy, time, and memory aspects altogether, we measure the per-image average clock time (in millisecond) in testing and as well conduct memory assessment by counting the number of trainable floating point parameters in the entire TDMSPC-Net. We summarize the results into the trade-off figure, (Fig. 4 in [47]), for comprehensive comparison. The following Fig. 10 illustrate the comparison. From the figure, the red star is significantly above the frontier,

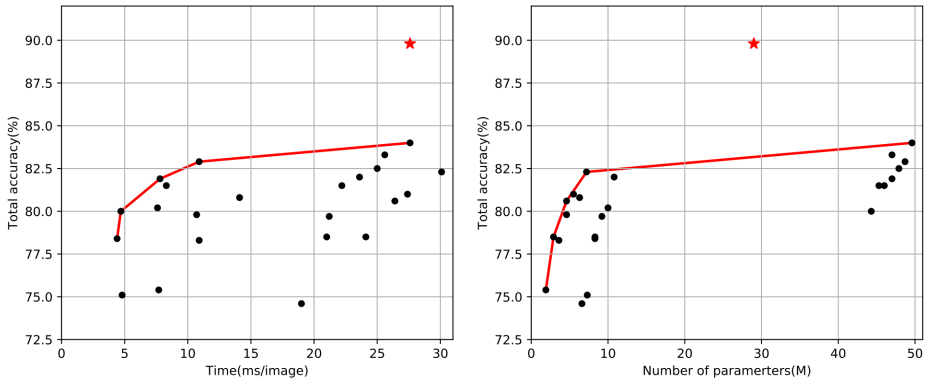

Figure 10: Accuracy versus time trade-off and accuracy versus memory trade-off of the proposed TDMSPC-Net. Black points represent the performances of combinations of all scene text recognition modules in [47]. Red solid curves indicate the trade-off frontiers found among scene text recognition combinations in [47]. Red star indicates the proposed TDMSPC-Net.

indicating that the proposed TDMSPC-Net achieve excellent accuracy-time and accuracy-memory trade off. This further proves the contribution of the proposed TDMSPC-Net.

References

[1] B. Su, S. Lu, Accurate scene text recognition based on recurrent neural network, in: Asian Conference on Computer Vision, Springer, 2014, pp. 35–48.

[2] B. Shi, X. Bai, C. 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) (2016) 2298–2304.

[3] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, Robust scene text recognition with automatic rectification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4168–4176.

[4] W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, J. Han, Star-net: a spatial atten-

tion residue network for scene text recognition., in: BMVC, Vol. 2, 2016, p. 7.

[5] C.-Y. Lee, S. Osindero, Recursive recurrent nets with attention modeling for ocr in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2231–2239.

[6] Y. Gao, Y. Chen, J. Wang, M. Tang, H. Lu, Reading scene text with fully convolutional sequence modeling, Neurocomputing 339 (2019) 161–170.

[7] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271.

[8] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, Convolutional sequence to sequence learning, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1243–1252.

[9] Y. Gao, Y. Chen, J. Wang, H. Lu, Reading scene text with attention convolutional sequence modeling, arXiv preprint arXiv:1709.04303.

[10] W. Liu, C. Chen, K.-Y. K. Wong, Safe: Scale aware feature encoder for scene text recognition, in: Asian Conference on Computer Vision, Springer, 2018, pp. 196–211.

[11] Z. Liu, Y. Li, F. Ren, W. L. Goh, H. Yu, Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[12] W. Liu, C. Chen, K.-Y. K. Wong, Char-net: A character-aware neural network for distorted scene text recognition, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[13] F. Zhan, S. Lu, Esir: End-to-end scene text recognition via iterative image rectification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2059–2068.

[14] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, X. Bai, Scene text recognition from two-dimensional perspective, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8714–8721.

[15] M. Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, X. Bai, Symmetry-constrained rectification network for scene text recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9147–9156.

[16] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, Aster: An attentional scene text recognizer with flexible rectification, IEEE transactions on pattern analysis and machine intelligence.

[17] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[18] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, S. Zhou, Focusing attention: Towards accurate text recognition in natural images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5076–5084.

[19] P. Wang, L. Yang, H. Li, Y. Deng, C. Shen, Y. Zhang, A simple and robust convolutional-attention network for irregular text recognition, arXiv preprint arXiv:1904.01375.

[20] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315–323.

[21] Y.-C. Wu, F. Yin, X.-Y. Zhang, L. Liu, C.-L. Liu, Scan: Sliding convolutional attention network for scene text recognition, arXiv preprint arXiv:1806.00578.

[22] F. Yin, Y.-C. Wu, X.-Y. Zhang, C.-L. Liu, Scene text recognition with sliding convolutional character models, arXiv preprint arXiv:1709.01727.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (4) (2017) 834–848.

[24] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 447–456.

[25] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[26] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122.

[27] H. Ding, X. Jiang, B. Shuai, A. Qun Liu, G. Wang, Context contrasted feature and gated multi-scale aggregation for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2393–2402.

[28] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[29] C. Luo, L. Jin, Z. Sun, Moran: A multi-object rectified attention network for scene text recognition, Pattern Recognition 90 (2019) 109–118.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[32] X. Chen, T. Wang, Y. Zhu, L. Jin, C. Luo, Adaptive embedding gate for attention-based scene text recognition, Neurocomputing.

[33] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Synthetic data and artificial neural networks for natural scene text recognition, arXiv preprint arXiv:1406.2227.

[34] A. Gupta, A. Vedaldi, A. Zisserman, Synthetic data for text localisation in natural images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2315–2324.

[35] A. Mishra, K. Alahari, C. Jawahar, Top-down and bottom-up cues for scene text recognition, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2687–2694.

[36] K. Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 1457–1464.

[37] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, et al., Icdar 2003 robust reading competitions: entries, results, and future directions, International Journal of Document Analysis and Recognition (IJDAR) 7 (2-3) (2005) 105–122.

[38] D. Karatzas, F. Shafait, Uchida, et al., Icdar 2013 robust reading competition, in: 2013 12th International Conference on Document Analysis and Recognition, IEEE, 2013, pp. 1484–1493.

[39] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. Chandrasekhar, S. Lu, et al., Icdar 2015 competition on robust reading, 2015, in: International Conference on Document Analysis and Recognition (ICDAR).

[40] T. Quy Phan, P. Shivakumara, S. Tian, C. Lim Tan, Recognizing text with perspective distortion in natural scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 569–576.

[41] A. Risnumawan, P. Shivakumara, C. S. Chan, C. L. Tan, A robust arbitrary text detection system for natural scene images, Expert Systems with Applications 41 (18) (2014) 8027–8048.

[42] C. Luo, Moran-v2, https://github.com/Canjie-Luo/MORAN_v2/, accessed March 4, 2019.

[43] W. Wang, E. Xie, P. Sun, W. Wang, L. Tian, C. Shen, P. Luo, Textsr: Content-aware text super-resolution guided by recognition, arXiv preprint arXiv:1909.07113.

[44] F. Bai, Z. Cheng, Y. Niu, S. Pu, S. Zhou, Edit probability for scene text recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1508–1516.

[45] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, S. Zhou, Aon: Towards arbitrarily-oriented text recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5571–5579.

[46] Y. Huang, Z. Sun, L. Jin, C. Luo, Epan: Effective parts attention network for scene text recognition, Neurocomputing.

[47] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, H. Lee, What is wrong with scene text recognition model comparisons? dataset and model analysis, arXiv preprint arXiv:1904.01906.