IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 6, 2021, accepted September 22, 2021, date of publication September 24, 2021,
date of current version October 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3115606

A Residual-Attention Offline Handwritten Chinese Text Recognition Based on Fully Convolutional Neural Networks

YINTONG WANG[1,2,3], YINGJIE YANG[3], WEIPING DING[4], AND SHUO LI[2,3]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
[2]School of Information Engineering, Nanjing Xiaozhuang University, Nanjing 211171, China
[3]Institute of Artificial Intelligence, De Montfort University, Leicester LE1 9BH, U.K.
[4]School of Information Science and Technology, Nantong University, Nantong 226019, China

Corresponding author: Shuo Li (shuo.li@dmu.ac.uk)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976118 and Grant 61806098, in part by the Joint Research Grant of Royal Society and National Science Foundation of China under Grant IEC\NSFC\170391, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180142, and in part by Jiangsu Government Scholarship for Overseas Studies under Grant JS-2019-104.

ABSTRACT Offline handwritten Chinese text recognition is one of the most challenging tasks in that it involves various writing styles, complex character-touching, and large number of character categories. In this paper, we propose a residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks, which is segmentation-free handwritten recognition that avoids the impact of incorrect character segmentation. By designing a smart residual attention gate block, our model can help to extract important features, and effectively implement the training of deep convolutional neural networks. Furthermore, we deploy an expansion factor to indicate the trade-off between computing resources for model training and the ability of a gradient to propagate across multiple layers, and make our model training adapt to different computing platforms. Experiments on the CASIA-HWDB and ICDAR-2013 competition dataset show that our method achieves a competitive performance on offline handwritten Chinese text recognition. On the CASIA-HWDB test set, the character-level accurate rate and correct rate achieve 97.32% and 97.90% respectively.

INDEX TERMS Offline handwritten recognition, convolutional neural networks, connectionist temporal classification, residual attention.

I. INTRODUCTION

Offline handwritten Chinese text recognition (OHCTR) is a challenging issue and has received significant attention from researchers [1]–[3]. The reason can be generally attributed to two important factors. The first one is the rapid growth of the OHCTR application requirements, including office handwritten document processing, mailing address recognition and precious historical manuscript recognition. The second one attributes to the inherent long-standing complexity of the OHCTR that involves various writing styles, complex character-touching, and large number of character categories. Although the development of neural networks has promoted the development of OHCTR to a certain extent, the existence of characters with high similarity and the significant difference in individual handwriting styles make it remain as an open problem.

The recognition of offline handwritten Chinese text usually adopts sequential pattern recognition techniques [4]–[6], which can be divided into two categories in general: over-segmentation approaches and segmentation-free approaches. The recognition approaches based on over-segmentation by integrating linguistic context model, character classifier, and topological geometric have been demonstrated to be successful in offline handwritten text processing. Both the character shape modeling and linguistic context model are playing very important role. They firstly acquire candidate segmentation-recognition topological geometric paths from consecutive over-segments, and then perform optimal path search by integrating linguistic context model and character classifier [7], [8]. There are some over-segmentation based systems combining with neural network language model to replace

The associate editor coordinating the review of this manuscript and approving it for publication was Qi Zhou.

the conventional character classifier, segmentation and geometric models, and they have achieved the best performance of over-segmentation based methods on the CASIA HWDB 2.0-2.2 datasets [9]. However, these over-segmentation methods are faced with the data sparseness problem which hinders these models from estimating optimal path stably as the number of parameters grows exponentially with the length of the topological path. Moreover, these over-segmentation methods are designed for specific situations and it is difficult for these methods to deal with the overlapping and touching characters in general.

The recognition approaches based on segmentation-free do not need to explicitly segment text lines into individual characters. One early approach to text line modeling using the Gaussian mixture model, hidden Markov model (HMM) is the representative which was firstly applied in speech recognition, and then introduced into the recognition of offline handwritten Chinese text lines and achieved good performance [4], [10], [11]. As the length of the recognition characters increases, HMM-based method involves excessive parameters, which leads to the degradation of its recognition performance. Different from these models, neural networks have completely changed many fields of machine intelligence, making the challenging task of image recognition successful with superhuman accuracy. One recent approach utilized recurrent neural networks(RNN) for the recognition of handwritten English languages with small number of character categories. The RNN approach is quite flexible and it avoids explicit segmentation which is largely due to the connectionist temporal classification (CTC) [12]. Suryani *et al.* [13] employed a CNN and LSTM under the HMM frame work to obtain a significant improvement over the traditional LSTM-HMM model. Wu *et al.* [14] used separable MDLSTM and RNN with CTC loss, instead of the traditional LSTM-CTC method. Although all those methods utilize recurrent architectures to properly conceal and learn serial context information, they lack the parallelization ability in the training phase and demand significant computing resource. Furthermore, handwritten text recognition is applied only in near neighbor character recognition or single character recognition to a large extent, and long-range dependencies have not been accounted so far.

For the above reasons, there is a recent shift to recurrence-free neural network architectures in most sequence recognition modeling works. We can see the trends of convolutional neural networks and connectionist temporal classification (CNN+CTC) in handwritten recognition problems [15], [16]. Peng *et al.* [17] proposed an end-to-end offline handwritten Chinese text recognition method using fully convolutional networks. This method is composed of three computation modules, include location, detection and classification, but its efficiency is not ideal. Liu *et al.* [18] proposed an efficient and effective offline handwritten text recognition method with convolutional neural networks for the challenging OHCTR task. This method performs convolution operations with equal weights on all input pixels, resulting in a decrease in the recognition performance of handwritten text with background noise. Mohamed *et al.* [19] proposed a novel handwritten recognition method based on fully convolutional neural networks. This method consists mostly of depthwise separable convolution operations with residual connections and softmax gating, trained on word or text line labels using the loss function of connectionist temporal classification. However, the method has very complex residual attention module and verifies its effectiveness only with small categories samples.

Inspired from the recent works on the convolutional neural networks [19]–[22], we proposed a residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks, which is segmentation-free handwritten recognition that avoids the impact of incorrect character segmentation. This method introduces a novel residual attention convolution to increase the importance of representative features and reduce the negative influences of the background or noise. More importantly, the expansion factor adjusts the number of tensor channels in the convolution process, balances the ability of a gradient to propagate across multiple layers. Our main contributions are as follows:

1) The recurrent-free architecture for offline handwritten Chinese text recognition is presented to avoid the degradation of recognition performance caused by character error segmentation and utilize the parallelization ability in the training phase and avoid the problem of large delay caused by recurrent iterative operation.
2) The novel smart residual attention gate block is designed to combine the advantages of residual framework and attention framework. Then the representative features are extracted and it alleviates the problems of gradient explosion or disappearance for deep convolutional neural networks.
3) The expansion factor is introduced to quantitatively analyze the performance of convolutional neural networks, so that the setting of the model parameters can achieves a balance between model performance and computing resources.
4) A set of extensive experiments are conducted on two widely adopted Chinese benchmark datasets: CASIA-HWDB and ICDAR-2013. The proposed method achieves competitive results in the benchmarks of character error rate without/with using language model, and demonstrates the validity of our proposed method.

The rest of the paper is organized as follows: Section 2 reviews briefly previous related works. In Section 3, we explain the proposed a residual-attention offline handwritten Chinese text recognition method. Experimental studies are given in Section 4. Finally, we draw the conclusions and future works in Section 5.

II. RELATED WORKS

In this section, some basic concepts of attention mechanism and connectionist temporal classification are discussed. Regarding the most mathematical symbols involved

in the work, we give a brief explanation here. Let $X = \{x_1, x_2, \cdots, x_n\}$ be an offline handwritten text line image set, where the i-th image $x_i \in \mathbb{R}^{h \times w \times 3}$. $L = \{l_1, l_2, \cdots, l_n\}$ represents the sequence label set of the image databases, where $l_i = \{l_{i1}, l_{i2}, \cdots, l_{in'}\}$, l_{ij} denotes the j-th character of the sequence label of the i-th image. $Y = \{y_1, y_2, \cdots, y_m\}$ is the tensor set, where $y_i \in \mathbb{R}^{h' \times w' \times c'}$ denotes the input tensor of the i-th layer of our model, h' is the height, w' is the width and c' is the number of channel.

A. ATTENTION MECHANISM

The attention mechanism can be considered as a kind of guidance which makes the allocation of available processing resources tend to the part with the most discriminate information in the input document image, and improve the problem of partial information loss caused by indispensable down-sampling, such as limited computing resource, network transmission and storage space [21]. Generally, it can reduce or even filter the influence of background noise of lower-resolution feature maps on the results, and strengthen the important parts in the input document image. Currently, some tentative researches have been proposed to combine the attention mechanism with deep neural networks [23], [24]. These researches range from positioning and understanding in complex images [25] to sequenced-based neural networks [26]–[28]. It is also usually integrated with a nonlinear activation function, Softmax or Dropout, to down-sampling or up-sampling the feature maps.

Each attention module can be divided into two calculation sub-modules, which are the trunk sub-module and mask sub-module. The trunk sub-module implements feature processing, and can be introduced into any advanced neural network structures. Given trunk sub-module output feature map $F_c(y_i)$ with input feature y_i, the mask sub-module utilizes a common bottom-up top-down calculation structure [29], [30] to obtain the same size output feature map $M_c(y_i)$, and it is used as the softly weight of output features map $F_c(y_i)$. In the mask sub-module, the design of bottom-up top-down calculation structure is derived from the fast feedforward and feedback process of neural networks. The output feature map of mask sub-module is employed to determine the nonlinear gates for neurons of the trunk sub-module. The output feature map of attention module $H_c(y_i)$ is:

$$H_c(y_i) = M_c(y_i) \times F_c(y_i). \tag{1}$$

B. CONNECTIONIST TEMPORAL CLASSIFICATION

Connectionist Temporal Classification (CTC) is a training criterion function, which can make an automatic alignment between the two sequences of the unknown input feature sequence and the known label sequence, devised for solving the sequence labeling problems [31], [32]. It has been widely used in many fields, such as human speech recognition [33], [34], handwritten recognition [1], [19], gesture language recognition [35] and continuous image segmentation and classification [36]. The CTC has proven to be effective in connectionist sequence recognition tasks.

Given the input feature sequence Y and the known label sequence L, CTC can learn the alignment without using the frame level alignment information, and its implementation requires four steps. Firstly, CTC deals with the repeated label in the known label sequence. For most Chinese text recognition tasks, CTC does not require the blank symbol to distinguish different phrases, but it is essential in English tasks. Secondly, it is employed as label for unlabeled feature sequence. It is worth noting that the nonlinear softmax function needs to normalize the outputs to get the distribution from l_t to Y at every timestep.

Thirdly, the complete output label sequence is employed to generate a distribution over all possible alignments, where each alignment is described as a possible path p_t. The path p_t is constituted of possible label sequence in L. We assume that the output label sequence of each timestep is independent of the other timesteps, the probability of one label sequence path p_t is:

$$p(\pi|Y) = \prod_{t=1:T} p(o^t_{\pi^t}|y^t). \tag{2}$$

where π^t indicates the label sequence at the timestep t for path π.

Finally, CTC then define a many-to-one mapping relationship F', it merges the first consecutive identical labels, and obtains the predicted label sequence of the label sequence path. In which, the probability of Label L can be processed as an optimization of the probabilities of all possible label sequence of the CTC. Its detail formula is as follows:

$$P(L|Y) = \sum_{\pi \in F'^{-1}(L)} p(\pi|Y). \tag{3}$$

As discussed, although much effort has been dedicated for offline handwritten Chinese text recognition methods based on convolution neural networks, the noted algorithms suffer from the following limitations and challenges.

1) Over-segmentation-based methods have the problem of data sparseness, in which they suffer from the unstable context estimation as the length of the context leads to the exponential number of parameters.
2) Most methods use recurrent architectures to correctly hide and learn successive sequence information, but they must confront the lack of parallelization during the training stage and require significant computing resource.
3) CNN-based methods are usually challenged by complex residual attention modules but their effectiveness has only been verified on some categories of small samples.

III. METHODOLOGY

In this section, we first introduce the framework of the residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks, and then give the concepts of residual attention gate block, which is

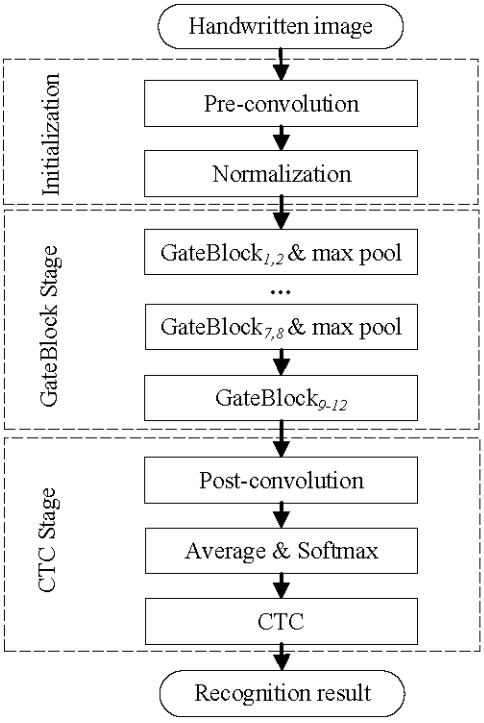


FIGURE 1. The flow chart of residual-attention offline handwritten Chinese text recognition.

the core computation of our method. Finally, we provide the model design and algorithm implementation.

A. FRAMEWORK

By introducing an attention mechanism and a residual network structure into convolutional neural networks, our method enhances the ability to extract meaningful features of the text line image and enhance the convergence ability of deep neural network training. More importantly, it only includes full convolutional operations, effectively avoids the demand on storage space and computing time due to loop or recursive operations. Therefore, our method can achieve a significant improvement in recognition accuracy and recognition efficiency.

Figure 1 shows the flowchart of residual-attention offline handwritten Chinese text recognition. There are mainly three stages:

(1) Initialization stage, the input text line image with $h \times w \times 3$ is processed with 1×1 convolutional operation to obtain a tensor with 31 channels, which is normalized with softmax to improve the convergence speed of neural network training. Then each channel is preprocessed independently with a 13×13 filter using a depthwise convolution operation, and the result of preprocessing is concatenated with the layer from normalized original image.

(2) GateBlock stage, as the core computational block of our method, it is composed of 12 GateBlocks. Each one is consisted of the attention mechanism and the residual structure, and adopts the separable convolutional operation to realize fast high-level feature extraction. Combining the model structure parameters in Table 1, there are four layers of the model, conv1.x-4.x, each layer includes two GateBlocks and a max pooling operation. The size of tensor is from $h \times w$ to $\frac{h}{8} \times \frac{w}{16}$, and the number of channels is from 32 to 1024. The conv5.x layer of the model contains four GateBlocks, and the size of tensors remains unchanged, all of which are $\frac{h}{8} \times \frac{w}{16} \times 1024$. Through the above GateBlocks, we can achieve the high-level representation abstraction of the input text line image.

(3) CTC stage, the two input sequences required by the CTC function are both one-dimensional, and we need to perform a post convolution operation on the tensor of size $\frac{h}{8} \times \frac{w}{16} \times 1024$ to obtain a tensor of size $\frac{h}{8} \times \frac{w}{16} \times C$, where C represents the number of character categories, and then apply the average pool and Softmax to this tensor on the h direction, obtain a prediction tensor of size $1 \times \frac{w}{16} \times C$. Then CTC is used to realize the alignment operation between the tensor and the label, the prediction result and the loss function value are obtained. In addition, in order to improve the convergence speed and normalize the tensor data in training process, we use multi-normalization methods in each layer [37, 38], such as batch normalization and layer normalization.

B. RESIDUAL ATTENTION GATE BLOCKS

Our method is constructed by stacking multiple GateBlocks as the main computational blocks. It uses attention gates to processing the flow of information between layers, and then increases the importance of representative features and reducing the importance of irrelevant features by weighting, this method has received significant interests from many researchers recently. In deep convolutional neural networks, attention mechanism is usually used together with residual structure to improve the convergence of the network. We opted to base the improved attention gates on the gating mechanism proposed by Deep Residual Learning [20]. Figure 2 shows the detailed structure of the i-th GateBlock.

Let $Y = \{y_1, y_2, \cdots, y_m\}$ be the output tensor set for each neural network layer, where the j-th tenor $y_i \in \mathbb{R}^{h' \times w' \times 2^k}$, where h', w' and k are determined by the corresponding network layer. Table 1 gives the parameters of our model. $H(\cdot)$ represents a mapping function from the input tensor to the output tensor of a network layer. That is, the i-th network layer's input tensor y_i, and its output tensor $y_{i+1} = H_c(y_i)$. In order to strive for the goal of the GateBlock, we propose a novel mapping function of residual attention gate block, whose formal expression is:

$$H_c(y_i) = [M_c(y_i) + 1] \times F_c(y_i) + y_i \tag{4}$$

where, $H_c(y_i)$ represents the c channel output sensor mapped from the input sensors respectively. $M(\cdot)$ represents a mask branch function, and $F(\cdot)$ represents a trunk branch function. In order to further understand the Eq.(4), we perform the identity transformation and get the new transformed expression is:

$$H_c(y_i) = M_c(y_i) \times F_c(y_i) + F_c(y_i) + y_i \tag{5}$$

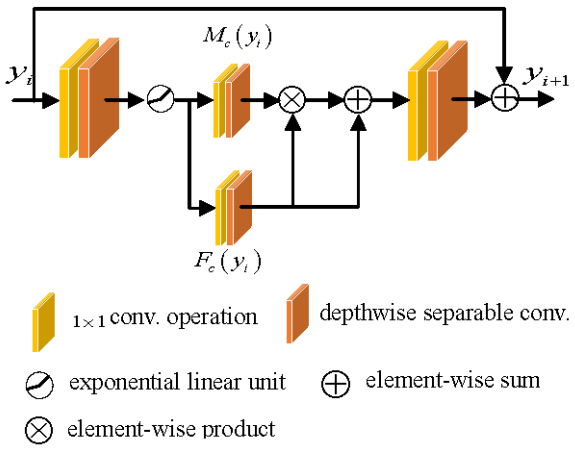


FIGURE 2. The detailed structure of the *i*-th GateBlock.

TABLE 1. The architectural details of the residual-attention offline handwritten Chinese text recognition.

Layer	Output size	Operator	n
	$h \times w \times 32$	1×1 Conv2d with 32 channels	
Conv1.x	$\frac{h}{2} \times \frac{w}{2} \times 128$	GateBlock	2
		2×2 max pool, stride 2	
Conv2.x	$\frac{h}{4} \times \frac{w}{4} \times 256$	GateBlock	2
		2×2 max pool, stride 2	
Conv3.x	$\frac{h}{8} \times \frac{w}{8} \times 512$	GateBlock	2
		2×2 max pool, stride 2	
Conv4.x	$\frac{h}{8} \times \frac{w}{16} \times 1024$	GateBlock	2
		2×2 max pool, stride 1×2	
Conv5.x	$\frac{h}{8} \times \frac{w}{16} \times 1024$	GateBlock	4
	$1 \times \frac{w}{16} \times C$	1×1 Conv2d with C categories	

In Eq.(5), let $H'_c(y_i) = M_c(y_i) \times F_c(y_i)$ and $H''_c(y_i) = F_c(y_i) + y_i$, and we get:

$$H_c(y_i) = H'_c(y_i) + H''_c(y_i) \tag{6}$$

The Eq.(6) is composed of two parts, where $H'_c(y_i)$ represents the attention module. For the i-th layer, the gate block from the input tensor y_i to the output tensor y_{i+1} contains two branches, $M_c(y_i)$ and $F_c(y_i)$. In which, $M_c(y_i)$ represents the mask branch of the attention module, and $F_c(y_i)$ represents the main branch of the attention module. They work together in the feature extraction process to encourage the representative feature information and depress the irrelevant feature information. In addition, $H''_c(y_i)$ indicates the residual network structure, it is also composed of two branches, $F_c(y_i)$ and y_i, respectively. In which, $F_c(y_i)$ represents the output tensor obtained by the mapping function or convolutional operation, and y_i is the output tensor directly transmitted to the next layer without any processing.

In order to make effective use highway gates even for wide and deep networks, we need to consider the dimensionality problem for the residual connection transformation. The $M_c(y_i)$ is firstly increased by 1, then multiplied by the convolution function $F_c(y_i)$, and finally added together with the input signal y_i. Here, we use the dual transformation mappings in Ref.[19]. Let P_1 be a negative transformation mapping $x \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$ to $x' \in \mathbb{R}^{\mathbb{H}'\times\mathbb{W}'\times\mathbb{C}'}$, and P_2 be a positive transformation mapping $x' \in \mathbb{R}^{\mathbb{H}'\times\mathbb{W}'\times\mathbb{C}'}$ to $x \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$. We can rewrite Eq.(4) such that

$$y'_i = P_1(y_i) \tag{7}$$

$$y_{i+1} = P_2\left([M_c(y'_i) + 1] \times F_c(y'_i)\right) + y_i \tag{8}$$

Using the negative transformation mapping P_1, it allows us to maintain the optimization benefit of residual connections whilst computing residual attention on different dimensional representations of y_i. P_1 and P_2 are implemented as depthwise separable convolutions, and the Exponential Linear Unit (ELU) as an activation function. In addition, we set $C' = tC$, $H' = H$, and $W' = W$. This means that down-sampling or up-sampling is only utilized on the channel dimension of the input data, and its spatial dimensions remain their original size. The expansion factor t is an exponential value with base 2, that is $t = [\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2]$. $t = 1$ means the number of tensor channels is constant; $t < 1$ represents the number of tensor channels is reduced, and GataBlock executes faster and uses less computational resource; $t > 1$ represents the number of tensor channels increases, and GataBlock has more available information and better feature representation ability.

C. MODEL DESIGN AND ALGORITHM IMPLEMENTATION

Our design is inspired by residual attention network [19], [21], which are widely used as backbone neural network and proven to have excellent performance for various tasks. The architecture of our method is depicted in Table 1, the first and last lines represent the pre-convolution operation and post-convolution operation, and the others represent 12 GateBlocks and some max pooling. The network consists of successive convolution layers with 3×3 kernel size and residual attention gate blocks, as depth-separable convolution. Each GateBlock layer is followed by batch normalization, layer normalization during training, and exponential linear unit be used as the non-linearity activation function, which own to its robustness when used with low-precision computation. The shape of line image is $h \times w \times 3$, where $h = 128$, means the height of the line image is uniformly processed to 128 pixels after text line preprocessing. Our model can adapt to the text line recognition with variant length.

Algorithm 1 gives the pseudo code of residual attention gate block, which includes four input parameters, such as tensor y_i, number of channel c, size of convolution kernel k, expansion factor t. In the implementation of the algorithm, it includes not only the necessary operation to realize the Eq.(4), but also three user-defined functions: forward depth separable convolution function *forwardConv*(·), tensor duplication function *nGates*(·) and backward depth separable

convolution function $blackwardConv(\cdot)$. Finally, the output tensor y_{i+1} is returned.

Algorithm 1 Pseudo Code of Residual Attention Gate Block

Input: tensor y_i, channel number c, kernel size k, expansion factor t

Output: tensor y_{i+1}

Def gateBlock(y_i, c, k, t):
  $y'_i = forwardConv(y_i, c, k, t)$
  $y''_i = torch.ELU(y'_i)$
  $x_0, x_1 = nGates(y''_i, c, k, \alpha)$
  $x_0 = torch.tanh\_(x_0)$
  $x_1 = torch.sigmoid(x_1)$
  $x = (x_0 + 1)x_1$
  $z_i = backwardConv(x, c, k, t)$
  $z'_i = torch.ELU(z_i)$
  $y_{i+1} = z'_i + y_i$
  return y_{i+1}

IV. EXPERIMENTAL RESULTS

To facilitate comparative studies against existing models in this field, we evaluated the performance of our handwritten Chinese text recognition method on two widely adopted datasets: a large database of offline Chinese handwriting called CASIA-HWDB [39] and a small dataset from the ICDAR 2013 Chinese handwriting recognition competition, abbreviated as ICDAR-2013 [40]. The method was implemented on the desktop computer of Intel Core i9-9900K 3.60GHz CPU, programming with python3.6 in PyCharm. While for training CNN models, we used NVIDIA RTX 2080ti GPUs for acceleration.

A. EXPERIMENTAL PREPARATION

In the experiment, we compared our method against four well know offline handwriting text recognition methods [8], [9], [11], [12], [14], [17]. These methods involve text recognition technologies such as traditional character over-segmentation, CNN and CNN-LSTM, and they have all shown their advantages in their respective aspects. The experiments are on two datasets, CASIA-HWDB and ICDAR-2013. Table 2 gives their detail information. As an unconstrained handwritten text database, the CASIA-HWDB is divided into training set and testing set. CASIA-HWDB represents the CASIA-HWDB including 3,118,447 isolated characters samples. ICDAR-2013 with 91,519 isolated characters samples. Each text line image is preprocessed into 128×2400 pixels, where '0' represents character handwriting, and '1' represents a blank background. When the height or width is less than its values, '1' is padded for the blank part.

Our model is trained by segmented-free pairs of text line and corresponding label sequence without any character/frame-level alignment. In each epoch, training examples are sampled from the training set without replacement. There are 90% samples of the training set from CASIA-HWDB for training the classifiers, and the remaining 10% samples is used to estimate the confidence parameter. Although neural networks show good performance at handwritten recognition, available training data is often not sufficient to capture handwriting variation widely from writer to writer, for which we introduce grid distortion method [43] to implement the augmentation of CASIA-HWDB database, shown as in Figure 4(b). We implemented our system on TensorFlow [44] deep learning framework, with the Adam [45] optimizer. All experiments use an initial learning rate of 5×10^{-3}, which is exponentially decayed to 0.001 after 1×10^6 batches; the maximum batch allowed by our platform (not less than 4, 2×2 max pool) is applied successively between layers until the height reaches 8 pixels. The architecture details of our method are shown in Table 1. In addition, the condition for model training to stop is that the maximum number of training times is 1 million iterations or the value of the loss function has not been improved for 50 consecutive iterations.

TABLE 2. The detail information of benchmark datasets.

		CASIA-HWDB	ICDAR-2013
Training	Pages	4076	—
	Lines	41781	—
	Characters	1081508	—
Testing	Pages	1015	300
	Lines	10449	3432
	Characters	267906	91519
Class		2703	2703

Levenstein edit distance [41] is used to measure the performance of the model on character level, and through the length of the label sequence to achieve normalization, which is commonly known as Character Error Rate (CER). In this paper, based on the literature [7], [9], [12], [14], the accurate rate (AR) and correct rate (CR) are employed to evaluate our model. Their formal expressions are as follows:

$$AR = \frac{N_t - D_e - I_e - S_e}{N_t} \tag{9}$$

$$CR = \frac{N_t - D_e - S_e}{N_t} \tag{10}$$

where N_t represents the length of the label sequence in the transcript. S_e, D_e and I_e represent the numbers of substitution errors, deletion errors and insertion errors, respectively.

B. RESULT AND DISCUSSION

1) EXPANSION FACTOR

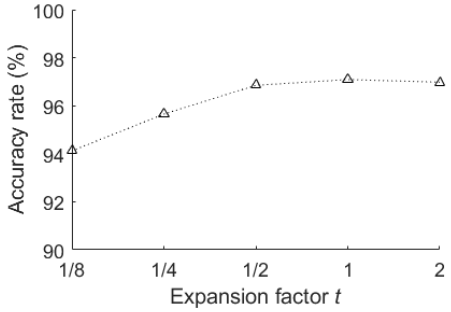
The expansion factor determines the number of channels in the GateBlock calculation process. It can down-sampling or up-sampling the original representation into low-dimensional or high-dimensional space, and performs lightweight depthwise convolution on the new representation, and then up-sampling or down-sampling the representation into the same size of the original dimensional space [42]. Table 3 gives the results with different expansion factors on the

TABLE 3. The recognition accuracy of two datasets with language model (%).

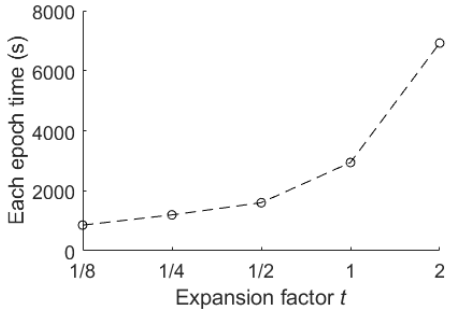
Indexs	Expansion factor				
	1/8	1/4	1/2	1	2
AR(%)	94.12	95.65	96.85	97.09	96.98
Time(s)	855	1196	1602	2941	6921
Size(MB)	47	66	115	258	702

CASIA-HWDB dataset, where the columns represent the expansion factor t, the three rows of the table indicate the index values obtained by our method on the CASIA-HWDB dataset, 'AR' is the accurate rate, 'Time' is the time of each epoch training, and 'Size' is the size of trained model. It can be known that the minimum and maximum of accurate rate are 94.12% and 97.09% respectively, and their variation range is 3.16%. However, the minimum and maximum of each epoch time are 855s and 6921s, the minimum and maximum of trained model size are 47MB and 702MB, and their increases have reached 809% and 1494% respectively.

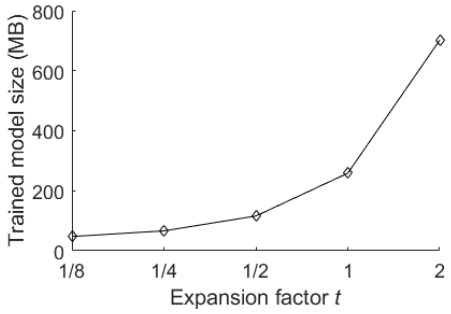
In addition, Figure 3 shows the trend of accurate rate, training time and trained model size of CASIA-HWDB test set under different expansion factors. The horizontal axis represents the expansion factor, and the vertical axis represents the accuracy rate, each epoch time and trained model size, respectively. As can be seen from Figure 3(a), the growth of accurate rate is relatively gentle and tends to converge to a certain value. On the contrary, the each epoch time and trained model size increases rapidly with the increase of the expansion factor in Figure 3(b) and 3(c). It is worth noting that reasonable setting of the expansion factor is important for handwritten recognition. When the expansion factor is set too small, the model can complete the training task in less computing space and time, but the accurate rate is not high; when the expansion factor is set too large, the trained model can obtain higher accurate rate, but more computing space and time are needed to complete the model training task. Through multi-experiments verification, we set the expansion factor $t = 1/2$ which can achieve a balance to a certain extent.



(a) Accuracy rate



(b) Each epoch time



(c) Trained model size

FIGURE 3. The trend analysis with different expansion factors on the CASIA-HWDB test set.

2) RECOGNITION ACCURACY

To verify the proposed module, we evaluate it against six state-of-the-art methods in the offline handwritten Chinese text recognition literature. These methods involve text recognition technologies such as traditional character over-segmentation, CNN and CNN-LSTM, as well as data augmentation or language model adaptation used to enhance the recognition accuracy. The recognition results on CASIA-HWDB and ICDAR-2013 datasets using compared methods are shown in Table 4 and Table 5. Each row represents a method and the best rates are indicated in bold face.

From Table 4, our method achieves the best recognition performance on the CASIA-HWDB without using any language model, where the AR and CR are 96.85% and 97.46%, respectively. Compared with the suboptimal method, the AR recognition performance of our method is improved by 2.05%. In the ICDAR-2013 dataset, our method obtains 91.30% for AR and 91.96% for CR. The recognition performance of our method is the competitive result among the seven methods in comparison. It is meaningful that our method, fully convolutional neural networks combining residual structure and attention mechanism, can achieve such results on offline handwritten Chinese texts involving variant writing styles, character-touching, and large number of character categories without the involvement of any language model.

To further increase the performance of our system, explicit language model is integrated to explore the semantic relationships between characters. By incorporating lexical constraints and prior knowledge about the language, language model can rectify some obvious semantic errors, thus improves the recognition result. In this paper, we only considered character tri-gram language model in experiments.

TABLE 4. The recognition accuracy of two datasets without language model (%).

Methods	CASIA-HWDB		ICDAR-2013	
	AR	CR	AR	CR
Du et al.[9]	—	—	83.89	—
NA-CNN[8]	92.04	93.24	88.79	90.67
SMDLSTM[14]	—	—	86.64	87.43
Wang et al.[11]	—	—	89.66	—
Peng et al.[17]	—	—	90.52	89.61
Xie et al.[12]	94.90	95.37	**91.55**	**92.13**
Ours	**96.85**	**97.46**	91.30	91.96

TABLE 5. The recognition accuracy of two datasets with language model (%).

Methods	CASIA-HWDB		ICDAR-2013	
	AR	CR	AR	CR
Du et al.[9]	—	—	93.50	—
NA-CNN[8]	95.21	96.28	94.02	95.53
SMDLSTM[14]	—	—	92.61	—
Wang et al.[11]	—	—	96.47	—
Peng et al.[17]	—	—	95.51	84.88
Xie et al.[12]	96.97	97.28	**96.72**	**96.99**
Ours	**97.32**	**97.90**	96.51	96.76

As shown in Table 5, our method uses the language model to obtain the maximum value of AR and CR of CASIA-HWDB test set, which are 97.32% and 97.90% respectively. Similarly, the suboptimal values of AR and CR obtained on ICDAR-2013 data set are 96.51% and 96.76% respectively. Although the language model can improve the recognition accuracy of the offline handwritten Chinese text recognition method, we should realize its two limitations. Firstly, the effectiveness of the language model is limited by the dataset itself. When the character relationship in the text mark of the dataset conforms to the language model modeling, it is helpful to improve the recognition accuracy; otherwise, it has little impact on the recognition accuracy, and even decreases. Secondly, the language model has a weaker optimization effect on model with higher recognition accuracy. Du *et al.* method gains the AR 83.89% without language model, and achieves the AR 93.50% with language model and the recognition performance has been improved by 11.46%. Correspondingly, our method only improves the recognition performance by 5.71% using the language model compared with its performance without the language model.

Through the further analysis of the CASIA-HWDB and ICDAR-2013 data sets, we also find that the datasets restrict the recognition accuracy of the compared methods to a certain extent. On the one hand, there are some incomplete or unclear characters in the offline handwritten Chinese text dataset, which also affects the text recognition result to a certain extent. On the other hand, although data augmentation has been done in the experiment processing, there are still many types of Chinese characters, large differences in writing styles and a huge gap between the amount of available data and the amount of data required for model training.

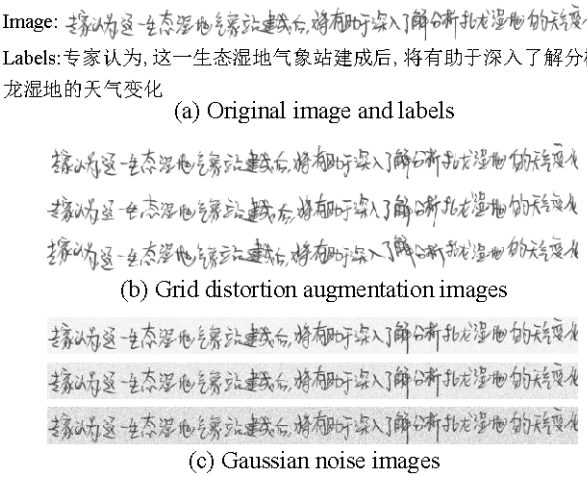


FIGURE 4. Example images of grid distortion augmentation and Gaussian noise.

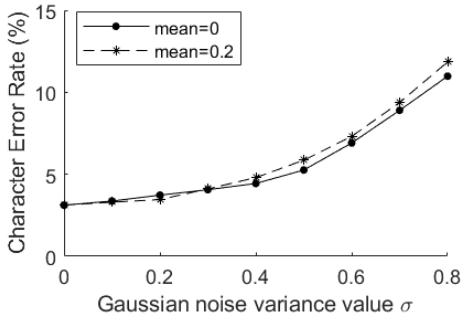


FIGURE 5. The trend analysis with different expansion factors on the CASIA-HWDB dataset.

3) MODEL ROBUSTNESS ANALYSIS

In order to verify the robustness of our model, the residual attention gate block can increase the importance of representative features (handwriting pixels) and reduce the importance of irrelevant features (background pixels) by weighting in the feature extraction process. We introduce grid distortion augmentation [43] and Gaussian noise on CASIA-HWDB database, as shown in Figure 4. We know that the character handwriting pixels are black (represented by '0'), and the background pixels are white (represented by '1'). Here, subfigure 4(a) is the original image and labels, and the size of image is 128 × 2400 pixels. Subfigure 4(b) shows the augmentation images with grid distortion, the interval is sets as {32, 64, 128} pixels and the variance sets as {6, 12, 24}. Subfigure 4(c) gives the Gaussian noise images with mean $m = 0$ and variance $\sigma \in [0.1 : 0.1 : 0.8]$. With the increase of σ, the amount of noise contained in the image increases, which has a greater challenge for text recognition.

It can be seen from Figure 5, we know that the character error rate of our model on the Gaussian noise database is

increased from 3.15% to 13.85%, with an increase range of less than 11%. Furthermore, when $\sigma \leq 0.3$ in Guassian noise, the CER of our method is still better than the compared methods. It shows that our model can realize the recognition of noisy offline handwritten texts to a certain extent, and has good robustness.

V. CONCLUSION

In this paper, we have proposed a residual-attention offline handwritten Chinese text recognition based on fully convolutional neural networks. A smart residual attention gate block has been designed to increase the importance of representative features and reduce the importance of irrelevant features by weighting, and it help to alleviate the problems of gradient explosion and gradient disappearance for deep convolutional neural networks. The expansion factor adjusts the number of tensor channels in the GateBlock convolution process, balances the computing resources for model training and the ability of a gradient to propagate across multiple layers. Experiments show that our method exhibits superior performance on CASIA-HWDB database.

In the future, we will continue to design and optimize the structure of the fully convolutional neural networks for offline handwritten Chinese text recognition, so that the method can be deployed under practical computational and other resource constraints.

REFERENCES

[1] Z. Chen, F. Yin, X.-Y. Zhang, Q. Yang, and C.-L. Liu, "MuLTReNets: Multilingual text recognition networks for simultaneous script identification and handwriting recognition," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107555.

[2] Z.-R. Wang, J. Du, and J.-M. Wang, "Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107102.

[3] Z. Xie, Y. Huang, L. Jin, Y. Liu, Y. Zhu, L. Gao, and X. Zhang, "Weakly supervised precise segmentation for historical document images," *Neurocomputing*, vol. 350, pp. 271–281, Jul. 2019.

[4] T.-H. Su, T.-W. Zhang, D.-J. Guan, and H.-J. Huang, "Off-line recognition of realistic Chinese handwriting using segmentation-free strategy," *Pattern Recognit.*, vol. 42, no. 1, pp. 167–182, Jan. 2009.

[5] Q.-F. Wang, F. Yin, and C.-L. Liu, "Unsupervised language model adaptation for handwritten Chinese text recognition," *Pattern Recognit.*, vol. 47, no. 3, pp. 1202–1216, Mar. 2014.

[6] Y. Wang, W. Xiao, and S. Li, "Offline handwritten text recognition using deep learning: A review," in *Proc. Int. Conf. Adv. Algorithms Control Eng.*, 2021, pp. 1–7.

[7] Q.-F. Wang, F. Yin, and C.-L. Liu, "Handwritten Chinese text recognition by integrating multiple contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1469–1481, Aug. 2012.

[8] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, and S. Naoi, "Deep knowledge training and heterogeneous CNN for handwritten Chinese text recognition," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 84–89.

[9] Y.-C. Wu, F. Yin, and C.-L. Liu, "Improving handwritten Chinese text recognition using neural network language models and convolutional neural network shape models," *Pattern Recognit.*, vol. 65, pp. 251–264, May 2017.

[10] J. Du, Z.-R. Wang, J.-F. Zhai, and J.-S. Hu, "Deep neural network based hidden Markov model for offline handwritten Chinese text recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3428–3433.

[11] Z.-R. Wang, J. Du, W.-C. Wang, J.-F. Zhai, and J.-S. Hu, "A comprehensive study of hybrid neural network hidden Markov model for offline handwritten Chinese text recognition," *Int. J. Document Anal. Recognit.*, vol. 21, no. 4, pp. 241–251, Dec. 2018.

[12] C. Xie, S. Lai, Q. Liao, and L. Jin, "High performance offline handwritten Chinese text recognition with a new data preprocessing and augmentation pipeline," in *Proc. Int. Workshop Document Anal. Syst.*, 2020, pp. 45–59.

[13] D. Suryani, P. Doetsch, and H. Ney, "On the benefits of convolutional neural network combinations in offline handwriting recognition," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 193–198.

[14] Y.-C. Wu, F. Yin, Z. Chen, and C.-L. Liu, "Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 79–84.

[15] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: http://arxiv.org/abs/1803.01271

[16] Y. Gao, Y. Chen, J. Wang, M. Tang, and H. Lu, "Reading scene text with fully convolutional sequence modeling," *Neurocomputing*, vol. 339, pp. 161–170, Apr. 2019.

[17] D. Peng, L. Jin, Y. Wu, Z. Wang, and M. Cai, "A fast and accurate fully convolutional network for end-to-end handwritten Chinese text segmentation and recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 25–30.

[18] B. Liu, X. Xu, and Y. Zhang, "Offline handwritten Chinese text recognition with convolutional neural networks," 2020, *arXiv:2006.15619*. [Online]. Available: http://arxiv.org/abs/2006.15619

[19] M. Yousef, K. F. Hussain, and U. S. Mohammed, "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107482.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[22] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," 2015, *arXiv:1507.06228*. [Online]. Available: http://arxiv.org/abs/1507.06228

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[24] Y. Jiang, H. Yao, C. Wu, and W. Liu, "A multi-scale residual attention network for retinal vessel segmentation," *Symmetry*, vol. 13, no. 1, p. 24, Dec. 2020.

[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 2017–2025.

[26] L. Kang, P. Riba, M. Villegas, A. Fornés, and M. Rusiñol, "Candidate fusion: Integrating language modelling into a sequence-to-sequence handwritten word recognition architecture," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107790.

[27] J. Sueiras, V. Ruiz, A. Sanchez, and J. F. Velez, "Offline continuous handwriting recognition using sequence to sequence neural networks," *Neurocomputing*, vol. 289, pp. 119–128, May 2018.

[28] J. Michael, R. Labahn, T. Gruning, and J. Zollner, "Evaluating sequence-to-sequence models for handwritten text recognition," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1286–1293.

[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[30] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

[31] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369–376.

[32] L. Chao, J. Chen, and W. Chu, "Variational connectionist temporal classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 460–476.

[33] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.

[34] E. Emiru, S. Xiong, Y. Li, A. Fesseha, and M. Diallo, "Improving Amharic speech recognition system using connectionist temporal classification with attention model and phoneme-based byte-pair-encodings," *Information*, vol. 12, pp. 1–22, Feb. 2021.
[35] N. Cihan Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10023–10033.
[36] D. Huang, F. Li, and J. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 137–153.
[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[38] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450
[39] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "CASIA online and offline Chinese handwriting databases," in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 37–41.
[40] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "ICDAR 2013 Chinese handwriting recognition competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1464–1470.
[41] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, 1966.
[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[43] C. Wigington, S. Stewart, B. Davis, B. Barrett, B. Price, and S. Cohen, "Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 639–645.
[44] M. Abadi, P. Barham, and J. Chen, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16, 2016, pp. 265–283.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

YINTONG WANG received the B.S. degree in computer application from Southeast University Chengxian College, Jiangsu, China, in 2009, the M.S. degree in computer software and theory from Changchun University of Technology, Jilin, China, in 2012, and the Ph.D. degree from the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China, in 2016. He worked as a Visiting Scholar with De Montfort University, U.K., in 2020. He is currently a Postdoctoral Researcher with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, and a Lecturer with the Key Laboratory of Trusted Cloud Computing and Big Data Analysis, Nanjing Xiaozhuang University. His research interests include handwritten text recognition, deep neural, and dimensionality reduction.

YINGJIE YANG received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees in engineering from Northeastern University, Shenyang, China, in 1987, 1990, and 1994, respectively, and the Ph.D. degree in computer science from Loughborough University, Loughborough, U.K., in 2008. He is currently a Professor of computational intelligence and the Deputy Director of the Institute of Artificial Intelligence, De Montfort University, Leicester, U.K. His research interests include grey systems, fuzzy sets, rough sets, neural networks and their applications to civil engineering, transportation, environmental engineering, and management science. He is a Senior Member of IEEE Systems, Man, and Cybernetics Society and a member of Rail Research U.K. Association. He is also the Executive President of the International Association on Grey Systems and Uncertainty Analysis, and the Co-Chair of IEEE SMC Technical Committee on Grey Systems.

WEIPING DING received the Ph.D. degree in computer science from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2013. From 2014 to 2015, he was a Postdoctoral Researcher with the Brain Research Center, National Chiao Tung University, Hsinchu, Taiwan. In 2016, he was a Visiting Scholar with National University of Singapore, Singapore. From 2017 to 2018, he was a Visiting Professor with the University of Technology Sydney, Ultimo, NSW, Australia. He is currently a Professor with the School of Information Science and Technology, Nantong University, Nantong, China. His research interests include deep neural networks, multimodal machine learning, granular data mining, and uncertainty modeling in big data, co-evolutionary algorithm, and medical images analysis. He also serves on the Editorial Advisory Board for *Knowledge-Based Systems* and the Editorial Board for *Information Fusion*, *Engineering Applications of Artificial Intelligence*, and *Applied Soft Computing*.

SHUO LI received the B.S. degree in computer science and technology from Nanjing University, China, in 2004, and the M.S. degree in computer application technology from East China Jiaotong University, China, in 2008. He is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence, De Montfort University, Leicester, U.K. He is also a Lecturer with the Key Laboratory of Trusted Cloud Computing and Big Data Analysis, Nanjing Xiaozhuang University. His research interests include pattern recognition and educational data mining.