

# Fast and Robust Online Handwritten Chinese Character Recognition With Deep Spatial and Contextual Information Fusion Network

Yunxin Li , Qian Yang, Qingcai Chen , *Member, IEEE*, Baotian Hu , Xiaolong Wang, Yuxin Ding, and Lin Ma 

**Abstract**—Deep convolutional neural networks have achieved fairly high accuracy for single online handwritten Chinese character recognition (SOLHCCR). However, in real application scenarios, users always write multiple characters to form a complete sentence, and previous contextual information holds significant potential for improving the accuracy, robustness and efficiency of recognition. In this work, we first propose a simple and straightforward model named the vanilla compositional network (VCN) by coupling convolutional neural network with a sequence modeling architecture (i.e., a recurrent neural network or Transformer), which exploits the handwritten character's previous contextual information. Although VCN performs much better than the previous state-of-the-art SOLHCCR models, it is a two-stage architecture in nature. It suffers from high fragility when confronting with poorly written characters such as sloppy writing, and missing or broken strokes, due to relying heavily on contextual information. To improve the robustness of the OLHCCR model, we further propose a novel deep spatial & contextual information fusion network (DSCIFN). It utilizes an autoregressive framework pre-trained on a large-scale sentence corpora as the backbone component, and highly integrates the spatial features of handwritten characters and their previous contextual information in a multi-layer fusion module. To verify the effectiveness of models, we reorganize a new form of online Chinese handwritten character with its previous context dataset, named OHCCC. Extensive experimental results demonstrate that DSCIFN achieves state-of-the-art performance and has increased strong robustness compared to VCN and previous SOLHCCR models. The in-depth empirical analysis and case study indicate that DSCIFN can significantly improve the efficiency of handwriting input because it does not need complete strokes to recognize a handwritten Chinese character precisely.

**Index Terms**—Multi-modal fusion, contextual information, Online handwritten Chinese character recognition (OLHCCR).

## I. INTRODUCTION

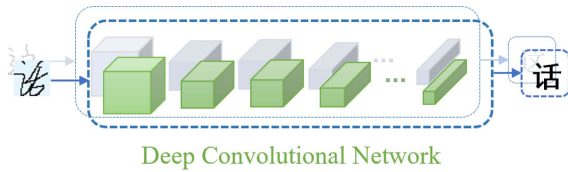
**H**ANDWRITTEN Chinese character recognition (HCCR) is an active research topic in the machine learning and pattern recognition field owing to its numerous applications, while it is a challenging task due to the problems posed by having a large number of classes, confusion of similar characters, and distinct handwriting styles. HCCR can be mainly divided into offline and online categories according to the acquisition of the used data [1]. Offline HCCR generally regards handwritten Chinese characters as gray-scaled or binary images, which are sensed by optical scanning or intelligent word recognition. In Online HCCR (OLHCCR), the strokes of each handwritten character can be obtained by picking up pen-tip movements as well as pen-up/pen-down switching. As the core of handwriting input method, OLHCCR finds various applications on pen input devices, personal digital assistants, smartphones, and computer-aided education. In this work, we focus on OLHCCR.

Deep convolutional neural networks (DCNNs) have advanced the state-of-the-art performance of OLHCCR. Most studies [1], [2], [4], [5] focus on single handwritten Chinese character recognition where each character is predicted independently without respect of its previous contextual information as shown in Fig. 1(a). In these works, the core challenge is how to model the details of the handwritten characters. There are two significant problems for these approaches in applications. On the one hand, the users are required to write the character completely, which limits the efficiency of the handwriting input. On the other hand, these models are relatively fragile when confronting sloppy writing, and missing or broken strokes of similar characters. As shown in Fig. 1(b), ‘话’ (talk) written by Person 1 (the last handwritten character) is extremely similar to the character ‘活’ (live) because of their sloppy writing of ‘讠’ and ‘讠’. Our preliminary investigation indicates that even the state-of-the-art model SS-DCNN [6], which exploits deep convolutional representations of stroke sequential information and eight well-designed directional features, fails to distinguish them correctly. In fact, users always write consecutive Chinese characters to form complete sentences in real applications. The contextual information holds significant potential to improve the accuracy, robustness, and efficiency of OLHCCR. According to the previous context in Fig. 1(b), it is not difficult to infer that the meaning of the sentence is “Do not like to talk to strangers”. Therefore, ‘话’ (talk) is more likely to be the correct word than ‘活’ (live). In this

Manuscript received 6 August 2021; revised 6 November 2021 and 16 December 2021; accepted 3 January 2022. Date of publication 14 January 2022; date of current version 7 June 2023. This work was supported in part by the Natural Science Foundation of China under Grant 62006061, in part by the Strategic Emerging Industry Development Special Funds of Shenzhen under Grant JCYJ20200109113441941, and in part by the Stable Support Program for Higher Education Institutions of Shenzhen under Grant GXWD20201230155427003-20200824155011001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. (*Corresponding author: Baotian Hu.*)

The authors are with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China, and also with Meituan, Beijing (e-mail: 19S051054@stu.hit.edu.cn; 20s051056@stu.hit.edu.cn; qingcai.chen@hit.edu.cn; hubaotian@hit.edu.cn; xlwangsz@hit.edu.cn; yxding@hit.edu.cn; forest.linma@gmail.com).

Digital Object Identifier 10.1109/TMM.2022.3143324



(a) Each handwritten character is predicted independently.

Do not like to talk to strangers

Truth:        不 爱 和 陌 生 人 说 话

Person 1:    不 爱 和 陌 生 人 说 话

Person 2:    不 爱 和 陌 生 人 说 话

Person 3:    不 爱 和 陌 生 人 说 话

(b) Handwritten characters have their contextual information.

Fig. 1. (a) Represents the single online handwritten character recognition method. (b) Represents some handwritten sentences and the blue colored characters are with missing or broken strokes.

work, we aim to realize fast and robust OLHCCR by exploiting previous context and handwriting of Chinese characters in an end-to-end architecture.

The handwriting of Chinese characters and their previous context are two totally different modalities. The previous context is the sequential Chinese character which contains complex semantic information and probabilistic dependency of an enormous set of characters, while handwriting is the individualized spatial representation of handwritten Chinese characters. The challenges of fully exploiting both of them in an end-to-end architecture lies in two aspects. One is how to efficiently and robustly integrate the two different information in a complementary way. Most of previous works use multi-stage pipeline methods [7], [8], which first use deep convolutional architecture to recognize single handwritten characters and obtain the candidate characters and then use the statistical language model to determine the final prediction. This approach suffers from the notorious error propagation problem and is always complicated and inefficient. Another challenge is to model the probabilistic dependency of the previous context and the current handwritten character, which requires the large-scale (generally more than 1 million sentences) online handwritten Chinese character with its previous context dataset. Unfortunately, to the best of our knowledge, there is no such a large off-the-shelf dataset. Even in the largest popular online Chinese sentence recognition dataset CASIA-OLHWDB (2.0, 2.1, 2.2) [9], there are only 24561 handwritten sentences which were produced by 1,019 writers on 5092 text pages. It is far from large enough to estimate the complex dependency probability between characters. Moreover, the data storage pattern in CASIA-OLHWDB is handwritten text pages, and its task mode is recognition of the entire handwritten Chinese sentence, which is different from our OLHCCR. Furthermore, manually hand-writing a large dataset to train a plausible end-to-end model is an extremely exhausting endeavor.

To address the technical challenge, we first propose a straightforward end-to-end network, namely vanilla compositional network (VCN). It couples a deep convolutional networks with sequence modeling structure such as recurrent neural networks [10]–[12] or transformer [13]. The VCN integrates previous contextual information based on the representation of each handwritten character yielded by DCNN. Although VCN incorporates the contextual information of handwritten characters, it can be regarded as a two-stage model where the handwriting and contextual information is encoded in two isolated modules. The character recognition relies on the probabilistic dependency of Chinese characters heavily and ignores the significance of handwriting details, which can easily lead to cascading errors. It exposes high fragility while confronting with poorly written characters. To integrate handwriting and contextual information in a complementary way, we further propose a more sophisticated model named deep spatial & contextual information fusion network (DSCIFN). DSCIFN contains the core multi-layer fusion module and a handwriting representation module. We adopt character embeddings to represent the previously recognized handwritten Chinese characters while predicting the current handwritten Chinese character at each time step, which overcomes the problem of inadequate representations of poorly written characters in VCN. More specifically, we use a transformer-based autoregressive framework as the backbone structure. The representation of handwritten Chinese characters obtained by the handwriting representation module and their contextual information are integrated multiple times in the multi-layer fusion module. Each layer of it contains the multi-head mask self-attention, multi-modal fusion, and multi-perceptron sub-layers in turn. The multi-modal fusion sub-layer is mainly used to fuse the handwriting information of handwritten characters and their contextual information obtained by the multi-head mask self-attention sub-layer. This proposed multi-layer fusion method that continuously integrates contextual information and spatial features of handwriting makes the DSCIFN stable and robust.

To address the lack of a large-scale online handwritten Chinese character with its previous context dataset, we first collect approximately 2.4 million sentences from the Chinese Wikipedia<sup>1</sup> and Sougou news corpus.<sup>2</sup> All of the characters of these sentences fall into the level-1 of GB2312-80,<sup>3</sup> which contains the commonly used 3755 Chinese characters. Then, for each character in level-1 of GB2312-80, we obtained 900 people's handwriting samples from public online single Chinese character recognition dataset CASIA-OLHWDB(1.0/1.1) [9], HIT-OR3C [14] and SCUT-COUCH2009 [15]. Finally, by replacing the characters of each sentence with handwriting samples, we construct a large-scale online handwritten Chinese character with its previous context dataset, named OHCCC, which will be openly released after this work is published. We conduct extensive experiments on OHCCC and the results demonstrate that DSCIFN significantly outperforms previous state-of-the-art

<sup>1</sup>[Online]. Available: <https://dumps.wikimedia.org/zhwiki/>

<sup>2</sup>[Online]. Available: [http://www.sogou.com/labs/resource/list\\_yuliao.php](http://www.sogou.com/labs/resource/list_yuliao.php)

<sup>3</sup>[Online]. Available: <http://www.sunchateau.com/free/fantizi/ziku/gbk.htm>

single OLHCCR models as well as VCN. The in-depth empirical analysis shows that DSCIFN is more robust and efficient than its competitors.

The main contributions of this work can be summarized as follows:

- 1) We present two end-to-end models for OLHCCR: the vanilla compositional network (VCN) and the deep spatial & contextual information fusion network (DSCIFN), which integrate the spatial features of handwritten characters and their previous contextual information.
- 2) To address the lack of a large-scale online handwritten character dataset with contextual information, we collect enormous sentences and handwritten character sets to construct a new form of Chinese online handwriting dataset, namely OHCCC.
- 3) Extensive experiments demonstrate that DSCIFN is more robust and efficient than VCN and previous state-of-the-art SOLHCCR models.

## II. RELATED WORKS

In this section, we introduce recent methods of OLHCCR and multi-modal fusion.

In early studies of handwritten Chinese character recognition, researchers focused solely on the structural features of characters. With the advent of statistical learning methods, most handwritten Chinese character recognition are divided into structural methods or statistical methods [16], [17], that mainly rely on the handcraft features. Significant progress has been made in the application of statistical methods based on handcrafted features. Recently, DCNNs such as ZFNet [18], GoogLeNet [19], VGG [20] and ResNet [21], have been shown to capture the complex features of images and perform well on many image tasks such as image classification [22], [23], mathematical expression recognition [24] and image representation [25], [26]. Hence, researchers have utilized DCNNs (e.g., MCDNN [27], SSD-CNN [6]) to learn multi-dimension features of handwritten Chinese characters, which achieves significant success [28]–[31]. The abovementioned traditional methods usually address issues in three stages: data processing, feature extraction, and classification [6].

For OLHCCR, data preprocessing usually removes the noises in data and normalizes the size of handwritten characters. General methods [32] convert each handwritten Chinese character into a 2-D image, and then normalized algorithms, e.g., bimoent normalization (BMN) [32], centroid-boundary alignment (CBA), and modified CBA (MCBA) [6] are used to handle the handwritten character image. Additionally, feature extraction step is at the core of OLHCCR because the distinguishable features directly affect the final classification result, which is the main research objective. Generally, directional and gradient features (image modeling) are jointly used by models to extract the features of handwritten Chinese characters. Bai and Huo [30] enriched the feature engineering of handwritten Chinese character recognition by transforming the four-direction features into eight-direction features, which was a great step forward for OLHCCR. Finally, classification is the last step of

OLHCCR. Support vector machines and fully connected neural networks have been widely used in various OLHCCR models and have achieved remarkable gains. However, the styles of people's Chinese character hand-writing are very different. Even when writing the same Chinese character, the handwriting of a person change over time, which produces a large challenge for the entire handwritten Chinese character recognition process.

In the previously proposed methods, the SSDCNN model proposed by Liu *et al.* [6] integrates the sequence of strokes and eight directional features of Chinese characters, achieving state-of-the-art performance on single handwritten Chinese character recognition. However, the application scenario of OLHCCR is generally at the sentence level, and previous methods ignore that the contextual information of handwritten characters has the potential to improve the accuracy and robustness of OLHCCR. Some previous sentence-level handwriting recognition models [7] confront the intrinsic difficulty of inferior efficiency due to the individualized handwriting and the complicated probability dependency of Chinese characters, compared to the sentence-level pinyin input method [33]. Especially for the traditional sentence-level OLHCCR based on statistical methods, it is a multi-stage network and faced the disaster of dimensionality. Recently, some studies [34]–[37] have proposed to use the language model LSTM to learn the contextual information of the handwritten text, demonstrating the effectiveness of contextual information for HCCR. However, there is no a large-scale handwritten Chinese character with its context dataset to help train a available end-to-end model. To advance the research of OLHCCR, we first reorganize a novel dataset by pairing millions of sentences from the large news corpora and hundreds of handwritten sets of single Chinese character to simulate an input scenario in which each handwritten character has associated contextual information.

The application scenarios of general OLHCCR contain the static handwriting features of handwritten characters as well as their contextual information. The input is generally a sentence composed of consecutive handwritten Chinese characters. It can be reduced to the sequential classification task according to the step-by-step inferring method. For similar natural language text generation tasks such as Abstractive Summarization [38], [39], Question Generation [40] and Paraphrase Generation [41], recurrent neural networks (RNNs) [10], [42] and Transformer-based models [43]–[47] can encode the natural language text and generate coherent text in a step-by-step manner. Hence, we apply a sequence modeling architecture (e.g., LSTM, Transformer) to capture the contextual information of handwritten characters and propose a simple vanilla compositional network (VCN) at first.

When confronting with not well written characters, the representation of handwritten Chinese characters obtained by DCNNs is insufficient. However, the character representation methods (e.g. one-hot encoding, distributed representation [48]) can convert any character into uniquely identified character embeddings without considering its spatial features [49]. Capitalizing on the remarkable progress in natural language processing brought by this method, we utilize character embeddings to represent previously recognized handwritten Chinese



characters and further propose a deep spatial & contextual information fusion network (DSCIFN). It integrates the contextual information and the extracted handwriting representation of sequential handwritten characters, containing two-modal features. Inspired by previous fusion methods of multi-modal representation such as Early Fusion LSTM [50], Tensor Fusion Network [51], Memory Fusion Network [52] and Low-rank Multi-Modal Fusion [53], we design a multi-modal fusion sub-layer in each block of DSCIFN to strongly integrate the cross-modal features.

### III. OUR METHODOLOGY

We first introduce the problem formulation of OLHCCR with the contextual information in Section III-A and introduce the details of Transformer in Section III-B. We propose the vanilla compositional framework in Section III-C. Finally, we introduce the more advanced deep spatial & contextual information fusion network in Section III-D.

#### A. Problem Formulation

We aim to achieve the fast and robust OLHCCR by exploiting previous context and handwriting of Chinese characters, which differs from previous works that predicted Chinese characters with no consideration of their previous context. Hence, the problem can be reformulated as follows. Given the current online handwritten Chinese character  $\mathbf{H}_i$  and its previous context  $\mathbf{C}_i$ , the corresponding Chinese character  $w_i$  of  $\mathbf{H}_i$  needs to be predicted.  $\mathbf{H}_i$  consists of multiple handwriting strokes, denoted as  $\mathbf{H}_i = (s_{i,1}, \dots, s_{i,j}, s_{i,k_i})$ , where  $s_{i,j}$  is the  $j$ th stroke and  $k_i$  is the number of strokes. The previous context  $\mathbf{C}_i$  contains  $i - 1$  previously predicted Chinese characters denoted as  $\mathbf{C}_i = (w_1, \dots, w_j, \dots, w_{i-1})$ , where  $w_j$  is the  $j$ th previous word.

#### B. Basic Transformer

Transformer [13] has achieved great success in many artificial intelligence fields [38], [40], [54]. It is an architecture that transforms one input sequence into another one, consisting of an encoder and a decoder. Compared to traditional sequence modeling structures such as LSTM [10], GRU [55], it only uses an attention-based mechanism and a fully connected neural network. The encoder and decoder are composed of modules that have the same structure, which consists mainly of multi-head attention and feed forward layers. Concretely, each token of the input sequence is first converted into a vector through the embedding matrix, and each token has a corresponding position vector to distinguish different words. For each layer of Transformer, the output of the previous layer initially passes a multi-head attention sub-layer, which can be calculated as shown in (1),

$$\begin{aligned} \text{MultiHeadAttention} &= \text{Concat}(\text{head}_1, \text{head}_2, \dots, \\ &\quad \text{head}_h) \mathbf{W}^h \\ \text{head}_i &= \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \end{aligned} \quad (1)$$

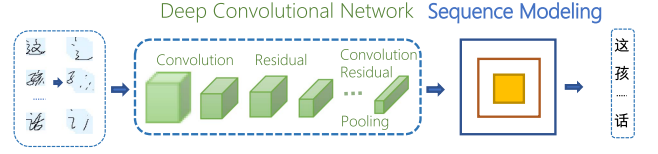


Fig. 2. The overall structure of vanilla compositional network composed of deep convolutional network and sequence modeling architecture.

where  $\mathbf{W}_i^Q$ ,  $\mathbf{W}_i^K$  and  $\mathbf{W}_i^V$  are the projection parameters, and their dimensions are  $\mathbb{R}^{d_{model} \times d_h}$ .  $\mathbf{W}^h \in \mathbb{R}^{hd_h \times d_{model}}$ , where  $d_{model}$  is the hidden size, and  $h$  is the number of heads. For each dot-product attention head, i.e., **Attention** [13], the computational method is denoted as given in (2).

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right) \mathbf{V} \quad (2)$$

Second, the output of the multi-head attention sub-layer passes the feed forward sub-layer, which is the fully connected neural network. Unlike the encoder, each layer of the decoder contains an additional multi-head attention sub-layer between the first multi-head attention sub-layer and the last feed forward sub-layer, which is used to interact with the output information of the encoder.

#### C. Vanilla Compositional Network

To exploit the previous context of the current handwritten character, a straightforward approach is to use a sequence modeling architecture such as recurrent neural network or transformer on top of the deep convolutional neural network, which is used to model the individualized handwriting of the current character. Therefore, we first propose a simple end-to-end vanilla combination network, named VCN, as shown in Fig. 2. VCN uses the stroke sequence-dependent deep convolutional neural network (SSDCNN-8) to capture the details of handwriting, which has been proven to be effective by Liu *et al.* [6]. Then, the fixed length vectorial representation in the final layer of SSDCNN-8 is fed into a sequence modeling architecture such as recurrent neural network or a transformer to estimate the probabilistic dependency of previous context and current handwritten character. Finally, a softmax layer is used to predict the target character of the input handwritten Chinese character.

#### D. Deep Spatial and Contextual Information Fusion Network

Although VCN exploits the previous context information and the spatial information of handwriting in an end-to-end fashion, it still is a multi-stage model in nature. VCN uses two isolated modules to model the handwriting of characters and the probabilistic dependency of Chinese characters respectively. It fails to fuse them together in a complementary way. In VCN, character recognition relies on the probabilistic dependency of Chinese characters heavily, while the spatial information of the handwriting is only used as the input of the sequence modeling module. This will lead to the diffusion of significant handwriting details and cascading errors when confronting with poorly written characters such as sloppy writing, and missing or broken strokes.

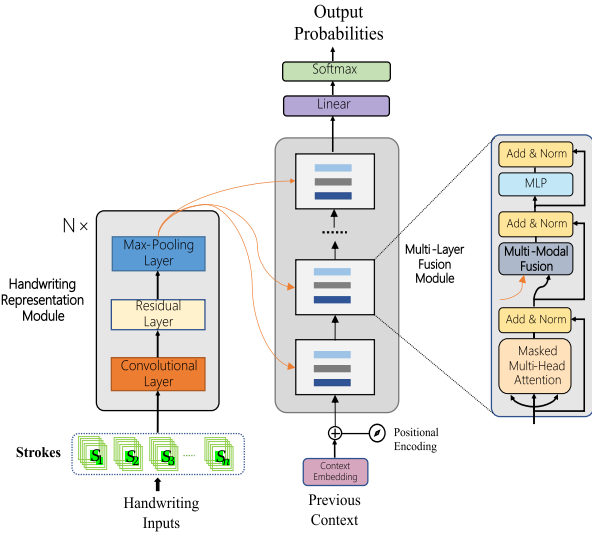


Fig. 3. Overview of the Deep Spatial & Contextual Information Fusion Network. The whole figure depicts a cyclic diagram of handwriting input. The input at each time step is the handwriting of current handwritten Chinese character and the handwritings and tokens of previously recognized handwritten characters.  $(S_1, S_2, \dots, S_n)$  represent the sequential multi-channel stroke maps of handwritten characters. The handwriting representation module is used to represent the handwriting input. The multi-modal fusion sub-layer is used to integrate the spatial feature of handwritten characters and their contextual information gained by the masked multi-head attention sub-layer. The first handwritten Chinese character has no previous information, so the decoder input is replaced by the special symbol *BOS*. The value of  $N$  is set to 4.

In view of the drawback of VCN, we propose the deep spatial & contextual information fusion network (DSCIFN), which is built on the deep interaction between the spatial information of handwritten characters and the sequential information of their previous context. As shown in Fig. 3, DSCIFN is composed of a *handwriting representation module* and a *multi-layer fusion module*. We introduce the structure and function of each module in detail below.

1) *Handwriting Representation Module*: The handwriting representation module is used to obtain the sequential representation of handwritten characters, which depends exclusively on the strokes of the Chinese characters. We use the aforementioned SSDCNN-8 as the handwriting representation framework except for the different output dimensions of the last fully connected layer as shown in the left part of Fig. 3. We set the image size of each stroke to  $32 \times 32$  and the maximum number of each character's strokes to 28. More importantly, for any handwritten Chinese character, we fill the gap between the last stroke and the maximum number set with full-zero feature maps. The position on each feature map is set to 1 in positions where the strokes pass; otherwise, the position is set to 0.

The architecture of handwriting representation module contains four identical blocks and each block has three sub-layers, which are the convolutional layer, residual layer and max-pooling layer in turn. The multi-channel stroke maps encoded by convolutional sub-layer of the  $l$ th layer is computed as denoted in (3).

$$z_c^l = \text{ReLU}(\mathbf{W}^l z^{l-1} + \mathbf{b}^l) \quad (3)$$

where  $z^{l-1}$  is the output of the  $l-1$ th layer.  $\mathbf{w}^l$  and  $\mathbf{b}^l$  are the parameters of the convolutional layer.  $\text{ReLU}$  [56] represents the activation function. The output  $z_c^l$  then passes the residual network layer [21] and the process is computed as denoted in (4).

$$z_r^l = \mathbf{W}_s^l z_c^l + \mathcal{F}(z_c^l, \mathbf{W}_r^l) \quad (4)$$

where the function  $\mathcal{F}(z_c^l, \mathbf{W}_r^l)$  represents the residual operation block. It generally contains a convolutional layer and an activation function. Note that we perform the projection parameter  $\mathbf{W}_s^l$  for the shortcut connection to ensure that the output and input of the residual block have matching dimensions.

The residual operation layer is followed by the max-pooling operation, which can expand the perception range and maintain the invariance of spatial rotation and translation. We utilize a  $2 \times 2$  window for multi-channel feature maps  $z_r^l$ , which is denoted as (5).

$$z^l = \text{Max}_{2 \times 2}(z_r^l) \quad (5)$$

The whole process is iterated through four convolutional blocks until the fixed length output vector is obtained. We denote the final representation of one handwritten Chinese character as  $\mathbf{R} = (r_1, \dots, r_m)$ , where  $m$  is the dimension of the output vector. Finally, the representation of sequential handwritten Chinese characters can be denoted as  $\mathcal{R} = (\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_n)$  where  $n$  is the length of the input sequence.

2) *Multi-Layer Fusion Module*: After obtaining the handwriting representation of handwritten characters, we propose the multi-layer fusion module as the right part in Fig. 3 to integrate the spatial feature of handwritten characters and their contextual information through a continuous interaction.

First, we project the previously recognized characters  $\mathbf{C}_i = (w_1, \dots, w_{i-1})$  into the fixed-dimensional character embeddings via a learnable embedding table, with the start symbol *BOS*. Meanwhile, we incorporate the learned positional information for the sequence as shown in Fig. 3 to help learn the probability dependency relationship between characters. After the character representation is obtained, we pre-train transformer-decoder-based autoregressive framework by the task of predicting the next word, where each layer contains two sub-layers, i.e. masked multi-head attention layer and multiple perceptron layer. Such pre-training method aims to learn the probability dependency between different characters in advance. Ideally, the pre-trained autoregressive framework could capture the dependency between characters and predict the following character accurately, given the previous context.

However, it is notable that consecutive handwriting input has continuous strokes input of the target characters at each time step of prediction besides the outputs of previous steps compared to text generation task [57]. Inspired by this observation, we incorporate the handwriting prompt information of the target character into each layer of the autoregressive framework by adding the extra multi-modal fusion sub-layer. The other sub-layers are identical as the structure of pre-trained autoregressive framework. The overall multi-layer fusion module has four identical blocks, as shown in Fig. 3. Different from VCN, this whole architecture integrates the spatial information of handwritten Chinese

characters as well as its contextual information in a complementary manner. Each block can be computed as given in (6).

$$\begin{aligned} \mathbf{h}_l^M &= \text{LayerNorm}(\mathbf{h}_{l-1} + \text{MultiHeadAttention}(\mathbf{h}_{l-1})) \\ \mathbf{h}_l^F &= \text{LayerNorm}(\mathbf{h}_l^M + \text{MultiModalFusion}(\mathbf{h}_l^M)) \\ \mathbf{h}_l &= \text{LayerNorm}(\mathbf{h}_l^F + \text{MLP}(\mathbf{h}_l^F)) \end{aligned} \quad (6)$$

where  $\mathbf{h}_{l-1}$  is the output of the  $l-1$ th layer. **MultiHeadAttention** is based on the multi-head attention mechanism, which can capture the contextual information of each handwritten character. Its computational process [13] has been presented earlier in (1).

After obtaining the previous contextual information of each handwritten character, we fed them into the **MultiModalFusion** to fuse with the spatial features of the target character, as denoted in (7).

$$\mathbf{y} = \mathbf{W}_F^l [\mathbf{h}_l^M, \mathcal{R}] + \mathbf{b}^l \quad (7)$$

where  $[\cdot]$  represents the concatenate operation and  $\mathbf{W}_F^l$ , and  $\mathbf{b}^l$  are the projection parameters. The concatenation operation [50] is generally useful for integrating pieces of information from two modalities. The final sub-layer is **MLP** [58], which is composed of the fully connected neural network and activation function as the feed-forward networks of Transformer [13]. The top output of the multi-layer fusion module, denoted as  $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_n)$ , goes through a linear and softmax layer [6], [59] to obtain the final prediction probability as provided in (8).

$$P(\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)) = \text{Softmax}(\mathbf{W}\mathbf{G} + \mathbf{b}) \quad (8)$$

where  $P(\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n))$  is the prediction probability of the whole sequence of input handwritten Chinese characters and  $\mathbf{W}$ ,  $\mathbf{b}$  are the related parameters.

To summarize, the architecture of DSCIFN mainly has two main advantages: 1) The process of employing character embeddings to represent recognized handwritten characters enriches their representation and further addresses the inadequate representation of poorly written characters. 2) The multi-layer fusion module fuses the contextual information of handwritten Chinese characters and their spatial features multiple times, highly exploiting the two-modal information. Therefore, the entire architecture of DSCIFN is more stable and stronger compared to VCN.

#### E. Training and Inference

Suppose the target sequence  $\mathbf{T} = (w_1, \dots, w_n)$  has been obtained, VCN can be trained by minimizing the loss function as given in (9).

$$\mathcal{L}_{sg} = - \sum_{i=1}^n \log P_i(x_i = w_i | s_{1,1}, \dots, s_{i,j}, \dots, s_{i,k_i}) \quad (9)$$

and DSCIFN can be trained via minimizing the loss function as (10).

$$\mathcal{L}_{DFTRN} = - \sum_{i=1}^n \log P_i(x_i = w_i | s_{1,1}, \dots, s_{i,j}, \dots, s_{i,k_i}; w_1, \dots, w_{i-1}) \quad (10)$$

TABLE I

THE DETAILED STATISTICS OF OHCCC. LEN-1/2/3(%) REPRESENT THE PROPORTION OF THE THREE LEVELS OF SENTENCE LENGTH, WHICH IS DIVIDED ACCORDING TO AN INTERVAL OF 10 WORDS

OHCCC	Size	Len-1(%)	Len-2(%)	Len-3(%)
# Train	2,386,485	31.63	34.58	32.79
# Dev	15,412	23.60	42.89	33.51
# Test	46,089	20.31	41.80	37.89

where  $(s_{1,1}, \dots, s_{i,j}, \dots, s_{i,k_i})$  represent the sequential handwritten Chinese characters and  $(w_1, \dots, w_{i-1})$  represent the target sequence of input handwritten characters before the  $i$ th position, i.e. the previous context of the  $i$ th character.

For inference, at each time step, the input of VCN is the sequential handwritten characters until the current time step, yet the input of DSCIFN has the characters recognized in the previous time steps in addition to the above inputs. Both of the outputs of the two methods are the recognition results of all handwritten Chinese characters up to the current time step.

#### IV. ONLINE CHINESE HANDWRITTEN CHARACTER WITH ITS CONTEXTUAL INFORMATION

##### A. Dataset Construction

1) *Handwritten Chinese Character Set*: We select 3755 commonly used characters as the Chinese character set for handwriting input based on the level-1 of GB2312-80, and further collect 900 sets of handwritten Chinese characters from the existing datasets such as CASIA-OLHWDB 1.0 and 1.1, HIT-OR3C and SCUT-COUCH2009, where only 513 sets of them contain the chosen 3755 Chinese characters. Among the 900 sets collected, We extract 200 complete sets of handwritten Chinese characters for testing and another 100 complete handwriting sets for validation set. The remaining 600 sets of handwritten characters are used for training.

2) *Sentence Corpora*: We collect a amount of texts from the Sougou news corpora and Chinese Wikipedia, and split them into sentences according to segmentation symbols. We clear up the special symbols in the sentences and filter out the sentences that include characters other than the chosen 3755 Chinese characters. We retain sentences that include between 10 and 40 words. The detailed statistics and the proportion of sentence lengths are reported in Table I, where LEN-1/2/3(%) indicate that the length of sentences is divided into three levels according to a 10 word interval.

3) *Reorganization*: We match the Chinese characters in every sentence with their corresponding handwritten forms to construct a new form of the Chinese online handwriting dataset named as OHCCC, where each handwritten character has its previous context information. Each Chinese character has approximately 900 handwritten samples, so each sentence can appear in millions of different varieties, two of which can be seen in Fig. 4. This simple yet effective method of data reorganization can simulate almost any online handwriting scenario based on a large-scale sentence corpus and various handwritten Chinese characters.



The child is introverted and does not like to talk to strangers  
 这孩子性格内向不爱和陌生人说话  
 Person 1: 这孩子性格内向不爱和陌生人说话  
 Person 2: 这孩子性格内向不爱和陌生人说话  
 Diversity: 这孩子性格内向不爱和陌生人说话

Fig. 4. Person 1/2 represent each character is from the same handwritten set. Diversity represents each character is from different handwritten sets.

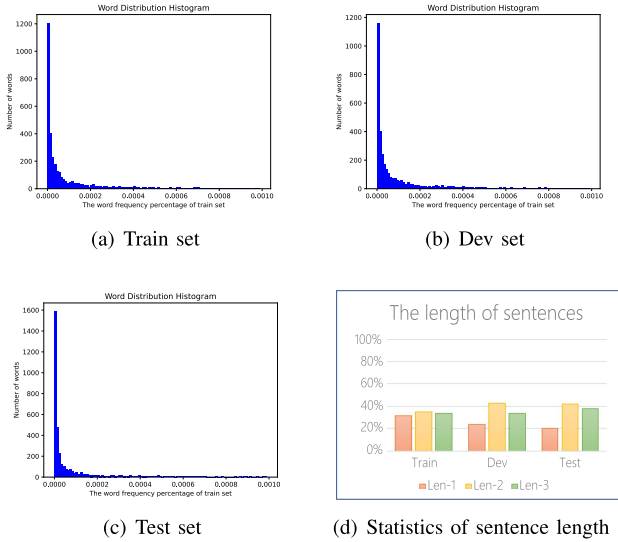


Fig. 5. The statistics of SOHCD, containing the word distribution of dataset and the distribution of sentence length.

## B. Data Statistics

1) *Data Analysis*: As shown in Table I, in order to better evaluate the effectiveness of models at sentence level, the training set includes a larger representation of sentences with more than 20 Chinese words compared to the uniform distribution in the validation and test sets. Note that all 3755 Chinese characters must be contained in the training, validation, and test sets to fully evaluate the chosen handwritten Chinese characters. Moreover, in Fig. 5(a) through Fig. 5(c), we observe that the word distributions of the training, validation and test sets are almost identical, which ensure the consistency of data features to a certain extent. Most words are low-frequency, while only a few commonly used words are high-frequency, which conforms to real human behavior when using Chinese.

## V. EXPERIMENTS

### A. Experiment Setting

1) *Models: Competitive SOLHCCR Models*: Various single OLHCCR models have achieved significant performance such as DSamCNN [60], DirMapCNN [17], DirMapCNN<sub>Adapt</sub> [17], SSDCNN [6], and SSDCNN<sub>Adapt</sub> [6]. They mostly used a DCNN to extract the handwriting feature of a single handwritten Chinese character, but sometimes they incorporate the additional handcrafted features and stroke sequence information as

well. The adaptation layer [61] in the above models aims to alleviate the recognition difficulty caused by the handwriting style of Chinese characters. It usually is added between the last fully connected layer and the softmax layer. We adopt the previous state-of-the-art SOLHCCR model SSDCNN, which achieves high performance and models the sequence of strokes as the main structure of our handwriting representation module. The stroke-dependent architecture (SSDCNN-8) is  $28 \times 32 \times 32$ -100C3ReLU-R3-MP2-100C2ReLU-R4-MP2-100C2ReLU-R4-MP2-200C2ReLU-R3-MP2-N256Sig. The eight-directional features are described by 512-N512Sig. The final classification layer of their combination can be denoted to 712-N300Sig-N200Sig-N3755.

a) *NET4*: Zhang *et al.* [34] proposed to utilize the recurrent neural network (RNN) LSTM/GRU to design the end-to-end OLHCCR model. It exploits a RNN to capture contextual stroke-dependent features, which achieve competitive performance compared to DCNN-based SOLHCCR models. In this work, we reimplemented the NET4 network they proposed and compared it with our model.

b) *VCN*: We select sequence modeling architectures LSTM and Transformer (abbreviated as xfmr) for the downstream network, named VCN (LSTM) and VCN (xfmr). We set the depths of LSTM and Transformer to 3 and 4, respectively. We change the dimension of the last fully connected layer to the hidden size of LSTM and Transformer. The simple compositional network effectively integrates the handwriting representation of Chinese characters and its contextual information through the language model.

c) *DSCIFN*: The size of the hidden state is 256 and the maximum sequence length is set to 512, identical as the aforementioned VCN model. Each block of it has eight attention heads. The architecture of multi-layer fusion module is  $512 \times 256$ -Transformer4-N6763. Compared to VCN, it not only enriches the representation of Chinese characters by introducing character embeddings, but also makes full use of the contextual information of Chinese characters.

2) *Implementation Details*: We construct our vocabulary on Chinese Internal Code Specification<sup>4</sup> with size 3755. For VCN and DSCIFN, the final output is 6763, yet it does not affect the final classification results. To speed up training, we initially load the stroke maps to memory instead of performing repeated loadings. Due to the limit of the memory size, we load 150 sets each time, and randomly reload the handwritten Chinese character sets after every 1000 steps to obtain new sets. For the selected 150 sets, we randomly select one handwritten sample for each character of the sentence. For VCN and DSCIFN, we pre-train LSTM, xfmr and Transformer-based autoregressive framework on the collected training sentences for 8 epochs to make them better understand the sentences and the maximum sequence length is set to 512. During training, each character in a sentence is represented by one  $28 \times 32 \times 32$  tensor and we train the above models on 3 V100 GPUs with the Adam optimizer with a base learning rate of  $1e-3$ , a batch size of 64, and a dropout rate of 0.1.

<sup>4</sup>[Online]. Available: <http://www.sunchateau.com/free/fantizi/ziku/gbk.htm>

TABLE II

COMPUTATIONAL COMPLEXITY ANALYSIS OF COMPETITIVE MODELS. 'MB (M)' REPRESENTS THE NUMBER OF PARAMETERS FOR EACH MODEL. 'TESTS' REPRESENTS THE TEST SPEED. FLOPS (MFLOPS) REPRESENT THE FLOATING POINT OPERATIONS PER SECOND

Model	MB (M)	Train Time (h/epoch)	TestS (chars/s)	FLOPs
DSamCNN	4.5	7.0	716	22
DirMapCNN	6.7	5.5	531	264
NET4	7.1	19.0	64	573
SSDCNN-8	5.7	15.4	138	709
SSDCNN	5.9	15.6	132	715
VCN (LSTM)	13.4	15.5	101	1067
VCN (xfmr)	11.2	16.2	86	902
DSCIFN	13.5	16.1	82	923
HRM	4.5	6.08	146	707
MFM	9.0	9.02	188	216

## B. Results and Analysis

In this section, we present and analyze the performances of VCN and DSCIFN introduced in Sections III-C and III-D on our designed experiments.

1) *Computational Performance*: The recognition speed of the OLHCCR model is important to its practical application. We tested the training and testing speeds of all models under the same hardware environment. To ensure the fairness of the test for the prediction speed of all models, we adopt 1000 sentences with the same handwriting style. As shown in Table II, we evaluate the train speed by counting the training time cost per epoch (i.e., Train Time (hours/epoch)). We also evaluate the test speed by counting the number of Chinese characters that each model recognizes per second (i.e., TestS (chars/s)). The greater the number of handwritten characters recognized, the faster its speed is considered be. 'HRM' and 'MFM' represent the handwriting representation module and multi-layer fusion module of DSCIFN, respectively. The DCNN models without considering the sequence of strokes achieve faster speed compared to those that incorporate the feature of the stroke sequence due to the small input feature dimension, e.g., DSamCNN vs SSDCNN: 716 vs 132 on test speed. The proposed VCN and DSCIFN models have slower training and testing speeds than other models, due to the larger number of parameters and more model layers. Moreover, we also analyze the computational complexity of the handwriting representation module and multi-layer fusion module of DSCIFN, and observe that their test speeds are fast. Hence, the proposed models could utilize parallel calculation when being used in practical applications.

2) *Accuracy*: We first test all models under the circumstance that each handwritten character has complete strokes, which can evaluate the best performance of models without considering the robustness and efficiency. During inference, the number of handwritten character sets is too large to be tested for each handwritten set. Hence, we randomly extract one handwritten sample from the 200 handwritten sets for each character and test **five** times for each model to ensure the confidence of experimental results. The average results are shown in Table III. Moreover, we select five sets from the total 200 test sets of handwritten

TABLE III

AUTOMATIC EVALUATION RESULTS ON THE TEST SET UNDER THE CONDITION THAT ALL HANDWRITTEN CHINESE CHARACTERS HAS THE FULL STROKES. P@I (I=1, 2, 3, 4, 5) REPRESENTS THE PROPORTION OF THE I RESULTS WITH THE HIGHEST PROBABILITY THAT CONTAINS THE CORRECT CATEGORY

Model	P@1	P@2	P@3	P@4	P@5
DSamCNN	93.83	95.62	96.78	97.73	97.74
DirMapCNN	95.65	97.87	98.60	98.90	99.08
DirMapCNN <sub>Adapt</sub>	95.74	97.99	98.60	98.92	99.08
NET4	96.15	97.67	98.15	98.39	98.53
SSDCNN-8	97.68	98.69	98.97	99.16	99.23
SSDCNN	97.74	98.71	99.05	99.21	99.35
SSDCNN <sub>Adapt</sub>	97.81	98.73	99.06	99.21	99.35
VCN (LSTM)	97.99	98.72	98.96	99.10	99.17
VCN (xfmr)	98.62	99.31	<b>99.49</b>	<b>99.55</b>	<b>99.61</b>
DSCIFN	<b>98.73</b>	<b>99.33</b>	99.45	99.53	99.59

TABLE IV

AUTOMATIC EVALUATION RESULTS ON THE TEST SET WITH FIVE SELECTED SETS OF HANDWRITTEN CHARACTERS. WE UNIFORMLY UTILIZE P@1 AS THE EVALUATION METRIC. HS-1, 2, ..., 5 REPRESENT THE SELECTED FIVE SETS OF SINGLE HANDWRITTEN CHINESE CHARACTER

Model	HS-1	HS-2	HS-3	HS-4	HS-5
DSamCNN	88.21	52.48	69.81	77.99	88.95
DirMapCNN	94.71	66.63	80.75	84.31	92.38
DirMapCNN <sub>Adapt</sub>	94.61	66.72	80.83	84.23	92.28
NET4	95.20	63.01	80.96	83.04	92.68
SSDCNN-8	95.26	70.20	81.14	90.03	97.05
SSDCNN	96.75	69.41	84.55	88.99	96.99
SSDCNN <sub>Adapt</sub>	96.71	69.81	84.56	88.91	97.01
VCN (LSTM)	97.04	71.56	88.45	91.43	97.03
VCN (xfmr)	97.87	76.46	89.29	91.80	98.25
DSCIFN	<b>98.15</b>	<b>78.47</b>	<b>92.25</b>	<b>91.89</b>	<b>98.32</b>

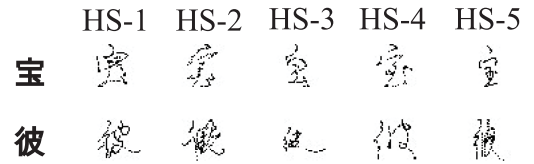


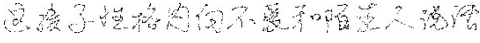
Fig. 6. Two handwritten Chinese character samples of the selected five sets. The written shapes of HS-2/3 are more sloppy than others.

Chinese characters and separately test character sets with the selected five forms for all models. The experimental result is shown in Table IV. We also select two handwritten character samples from the previously selected five handwritten sets, shown in Fig. 6, where the handwritten characters of HS-2/3/4 are extremely irregular. The five selected handwritten sets are also used in the following experiments. The experimental results in Table III show that VCN (xfmr) and DSCIFN perform well on all evaluation metrics, especially on P@1 where they both exceed SSDCNN approximately by 1 point. Compared to models that disregard the sequence of strokes (e.g., DirMapCNN, DirSamCNN), models that consider the sequence of strokes (e.g., NET4, SSDCNN) perform better on most evaluation metrics. From the experimental results of Table IV, we observe that the recognition accuracy of DSCIFN is much higher than that of SOLHCCR models especially on irregular and scribbled handwritten Chinese characters. The above experimental results demonstrate that SOLHCCR model can not recognize irregular and



The child is introverted and does not like to talk to strangers

这孩子性格内向不爱和陌生人说话

Full-Person A: 

DirMapCNN: 这孩子性格 **凶** 向不爱和陌 **半** 人说 **活**

NET4: 这孩子性格内向不爱和陌生 **久** 说 **活**

SSDCNN-8: 这孩子性 **红** 内向不爱和陌 **适** 人说 **活**

SSDCNN: 这孩子性格内向不爱和陌 **半** 人说 **活**

VCN (xfmr): 这孩子性格内向不爱和陌生人说话

DSCIFN: 这孩子性格内向不爱和陌生人说话

Fig. 7. Full-Person A represents one handwritten set with full strokes. Red colored words are the wrong recognition result (same as the figure below).

TABLE V  
AUTOMATIC EVALUATION RESULTS ON THE TEST SET UNDER THE CIRCUMSTANCE THAT EACH HANDWRITTEN CHINESE CHARACTER MISSES THE LAST STROKE

Model	P@1	P@2	P@3	P@4	P@5
DSamCNN	79.59	86.44	89.10	90.55	91.54
DirMapCNN	86.88	92.16	93.38	94.36	95.07
DirMapCNN <sub>Adapt</sub>	86.87	92.15	93.40	94.37	95.07
NET4	88.60	92.47	93.42	94.42	95.03
SSDCNN-8	88.87	92.37	93.62	94.35	94.85
SSDCNN	90.15	93.25	94.24	95.16	95.96
SSDCNN <sub>Adapt</sub>	90.12	93.22	94.27	95.22	95.96
VCN (xfmr)	90.27	93.02	93.95	94.50	94.89
DSCIFN	<b>92.81</b>	<b>95.22</b>	<b>96.01</b>	<b>96.45</b>	<b>96.72</b>

scribbled handwritten Chinese characters well without considering their context information, and models that incorporate the contextual information of handwritten characters are more robust and generalized. Compared to VCN (LSTM), VCN (xfmr) performs better on all evaluation metrics, which indicates that Transformer-based sequence modeling architecture with larger parameters can learn more precise probability dependencies between Chinese characters. An interesting phenomenon shown in Table III and Fig. 7 is that the recognition accuracy gap between DSCIFN and VCN (xfmr) is not great, which may be attributed to the fact that the handwriting representation of handwritten Chinese characters and their corresponding character embeddings have almost identical effect as their representation with complete strokes. It can inspire us to consider whether the glyph representation of Chinese characters can replace the current manner of character embeddings as a new representation method. However, for irregular handwriting such as HS-2/3, the performance of DSCIFN is significantly better than that of VCN.

3) *Robustness and Efficiency*: To evaluate the robustness of models and adapt to complex handwriting input scenarios, we test some strong competitive models under the scenario where each handwritten Chinese character retains only 70% strokes or losses its last stroke. We retrain the models introduced in Section V-A under the above cases of missing strokes, and the test procedure is identical as in cases of complete strokes except that the stroke retention of handwritten characters is different. The experimental results in Tables V to VIII show that DSCIFN significantly exceeds VCN and other SOLHCCR models on all evaluation metrics (e.g. P@1, HS-2/3). From the

TABLE VI  
P@1 ON THE TEST SET WITH FIVE SELECTED HANDWRITTEN SETS WHERE EACH HANDWRITTEN CHINESE CHARACTER MISSES THE LAST STROKE

Model	HS-1	HS-2	HS-3	HS-4	HS-5
DSamCNN	63.73	17.62	37.42	50.37	65.84
DirMapCNN	75.88	18.03	48.55	59.38	77.84
DirMapCNN <sub>Adapt</sub>	75.86	18.13	48.85	59.08	77.24
NET4	75.98	25.64	52.55	60.44	78.42
SSDCNN-8	75.34	31.10	55.94	58.95	78.27
SSDCNN	77.44	35.66	56.14	63.45	80.46
SSDCNN <sub>Adapt</sub>	77.54	35.45	56.73	63.71	80.34
VCN (xfmr)	78.67	35.48	56.82	63.41	79.73
DSCIFN	<b>81.59</b>	<b>37.98</b>	<b>60.95</b>	<b>70.45</b>	<b>82.00</b>

TABLE VII  
AUTOMATIC EVALUATION ON THE TEST SET UNDER THE CONDITION THAT ALL HANDWRITTEN CHINESE CHARACTERS RETAIN 70% OF STROKES

Model	P@1	P@2	P@3	P@4	P@5
SSDCNN	72.31	79.04	82.41	84.30	85.71
VCN (xfmr)	76.10	81.21	83.57	85.09	86.07
DSCIFN	<b>81.97</b>	<b>87.28</b>	<b>89.48</b>	<b>90.71</b>	<b>91.51</b>


TABLE VIII  
AUTOMATIC EVALUATION (P@1) RESULTS ON THE TEST SET WITH FIVE SELECTED HANDWRITTEN SETS WHERE EACH HANDWRITTEN CHINESE CHARACTER RETAINS 70% OF STROKES

Model	HS-1	HS-2	HS-3	HS-4	HS-5
SSDCNN	49.91	24.43	42.30	41.82	56.93
VCN (xfmr)	54.46	29.02	46.76	48.45	62.81
DSCIFN	<b>64.56</b>	<b>31.43</b>	<b>55.99</b>	<b>51.18</b>	<b>67.38</b>

Tables III to VI, the adaptation layer still slightly improves the performance of the corresponding basic model for each OLHCC in the case of missing strokes, and the performance of the models that consider the stroke sequence (e.g., NET4, SSDCNN, SSDCNN-8) is significantly better than the model that does not consider it. This indicates that the robustness of a model considering the sequence of strokes is better. The results further demonstrate that DSCIFN is more robust than other models because it experience less performance degradation. In addition, we observe that the recognition ability of DSCIFN decreases less than that of the other two models (i.e., SSDCNN, and VCN (xfmr)) as the strokes of handwritten characters decrease, especially on P@5. This phenomenon indicates that it is a feasible and robust method to integrate the handwriting features of handwritten characters and their contextual information multiple times based on the strong memory ability of pre-trained autoregressive framework. To compare the performance among models objectively, we take two examples from the test set and further analyse them. As shown in Fig. 8, some Chinese characters with similar structures becomes identical when missing strokes, such as the third word of the sentence form the end: ‘七’ (seven), ‘么’ (me) and ‘人’ (people) share the similar stroke ‘丿’. Hence, the spatial features of handwritten characters with missing strokes extracted by the deep convolutional network may easily become indistinguishable. This may cause the weak robustness of SSDCNN and VCN (xfmr). In addition, VCN (xfmr) that relies on contextual information heavily, does not integrate the handwriting prompts deeply, leading to cascaded errors. From Fig. 9, we

The child is introverted and does not like to talk to strangers

这孩子性格内向不爱和陌生人说话

MLS-Person A: 

SSDCNN: 这孩子性格内向 石乳和阵亡七说话


VCN (xfmr): 这孩子性格内向 不受和防生么说话

DSCIFN: 这孩子性格内向 不爱和陌生人说话

Fig. 8. MLS-Person A represents that each handwritten character is with missing the last stroke (same as the figure below).

His humorous, sincere and frank style has conquered the picky American business elites

他幽默风趣真诚坦率 的风格征服了挑剔的美国商界精英

70%-Person B: 

SSDCNN: 他幽默 凡趣真诚 坤率的 凡格征肥了挑剔的 美国商界精英

VCN (xfmr): 他幽默 凡趣真诚 坤率的 凡格征服了挑剔的 美国商界精英

DSCIFN: 他幽默 风趣真诚坦率 的风格征服了挑剔的 美国商界精英

Fig. 9. The second recognition instance depicts the prediction results under the circumstance that each handwritten character is with retaining 70% of all strokes.

observe that DSCIFN still recognizes ‘坦’ (smooth) accurately while it retains 70% of strokes, although it has a similar structure with another Chinese character ‘坤’ (Kun) at this time. This may be attributed to the fact that the phrase ‘真诚坦率’ (sincere and frank) is used often and DSCIFN can learn and utilize word combination information and handwriting information well.

From the above experimental results and instances, we observe that DSCIFN performs the best among all compared models for both simple/complex Chinese characters and complete/incomplete handwritten forms. In conclusion, DSCIFN is the most robust model among the compared models for recognizing online handwritten input. The handwriting method which uses the sequence modeling architecture combined with the contextual information of handwritten Chinese characters has a significant effect in improving its performance.

Improving the recognition accuracy of incompletely handwritten Chinese characters can promote the efficiency of online handwriting. Apart from the accuracy and robustness of the proposed models in the above experimental cases, we now analyze the efficiency of models when the strokes of handwritten Chinese characters are continuously reduced in sentences. We carefully evaluate the efficiency of the proposed models from two aspects. First is the overall recognition accuracy of the test set when the strokes of handwritten Chinese characters continuously decrease with an increase in the length of the sentence. Second is how many strokes models use at least to accurately recognize the position in a sentence.

First, we train the models on the training data in the scenario where handwritten Chinese characters in different positions retain different numbers of strokes with different probabilities. That is, characters in the first 25% of the sentence retain their strokes completely or only miss the last stroke; whereas the Chinese characters 25% ~ 50% of the way through the sentence retain 70% of their strokes; characters 50% ~ 75% of the way through the sentence retain only 50% of strokes; and Chinese characters in the last part of the sentence retain only 30% of

TABLE IX  
DIFFERENT POSITIONS RETAIN DIFFERENT PROPORTIONS OF STOKES DURING THE FIRST TRAINING PERIOD. 70%, 50%, AND 30% REPRESENT RETAINING THE CORRESPONDING PROPORTION OF STOKES FOR EACH HANDWRITTEN CHINESE CHARACTER

Retaining Position	Full	MLS	70%	50%	30%
0 ~ 25%	60%	30%	10%	0	0
25% ~ 50%	10%	40%	40%	10%	0
50% ~ 75%	10%	10%	30%	40%	10%
75% ~ 100%	10%	10%	20%	40%	20%

TABLE X  
CHARACTERS AT DIFFERENT POSITIONS RETAIN DIFFERENT PROPORTIONS OF STOKES DURING THE SECOND TRAINING PERIOD

Retaining Position	Full	MLS	70%	50%	30%
0 ~ 25%	5%	45%	50%	0	0
25% ~ 50%	0	30%	40%	30%	0
50% ~ 75%	0	10%	20%	40%	30%
75% ~ 100%	0	0	20%	30%	50%

TABLE XI  
AUTOMATIC EVALUATION RESULTS ON THE TEST SET UNDER THE CIRCUMSTANCE THAT STOKES OF HANDWRITTEN CHINESE CHARACTERS IN ONE SENTENCE ARE REDUCED AS THE LENGTH OF THE SENTENCE

Model	P@1	P@2	P@3	P@4	P@5
SSDCNN	71.80	80.86	85.58	87.23	88.55
VCN (xfmr)	84.07	90.79	93.13	94.33	95.22
DSCIFN	<b>90.68</b>	<b>94.61</b>	<b>95.93</b>	<b>96.66</b>	<b>97.14</b>

TABLE XII  
AUTOMATIC EVALUATION (P@1) RESULTS ON THE TEST SET WITH FIVE SELECTED HANDWRITTEN SETS

Model	HS-1	HS-2	HS-3	HS-4	HS-5
SSDCNN	67.78	41.33	54.35	56.21	66.06
VCN (xfmr)	78.40	51.64	66.21	68.54	78.53
DSCIFN	<b>85.41</b>	<b>57.29</b>	<b>73.78</b>	<b>79.66</b>	<b>86.77</b>

their strokes. More detailed experimental parameter settings are shown in Table IX. We first enable the model to learn with relatively complete strokes by retaining more strokes at different positions and then let the model learn by retaining fewer strokes, as shown in Table X. This makes the models move from learning from simple examples to more difficult ones over the course of learning. This training pattern allows models to learn how to fuse the spatial information and contextual information when the strokes are partially missing. During inference, for the first 75% of the sentence, we remove one stroke of the handwritten Chinese character according to 25% intervals and retain 70% of the strokes for the characters in the last 25% of the sentence. We test the models on the test set five times, and the average results are shown in Table XI The recognition accuracy of the selected five sets is shown in Table XII.

The experimental results of Tables XI and XII show that the performance of all models drop drastically on P@1, compared to the case of complete strokes. Yet DSCIFN can achieve approximately 97% accuracy on P@5. This shows that DSCIFN can maintain high recognition accuracy for each character as

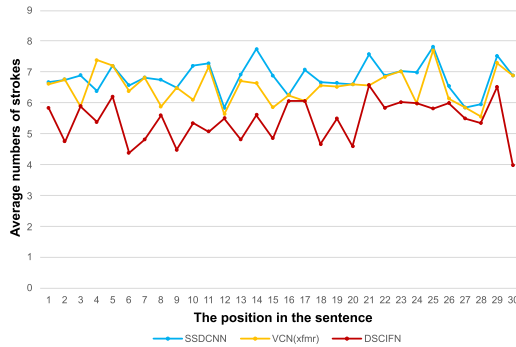


Fig. 10. The statistics of minimum number of strokes at each position recognized accurately by models. The evaluation metric of accuracy is P@3.

His humorous, sincere and frank style has conquered the picky American business elites  
 他幽默风趣真诚坦率的风格征服了挑剔的美国商界精英  
 Full Strokes: 他幽默风趣真诚坦率的风格征服了挑剔的美国商界精英  
 MNOS For DSCIFN: 个编甲几书下市书个几一初 划口 另 声四半 一

Fig. 11. An instance of correct recognition under the minimum number of strokes. MNOS FOR DSCIFN is the minimum number of strokes for DSCIFN.

the strokes of Chinese characters are reduced one at a time. The above phenomenon indicates that DSCIFN can promote the efficiency of online handwriting input.

The high recognition accuracy of p@1 can further promote the overall accuracy and efficiency of inferring sentence-level handwriting due to the intrinsic step-by-step prediction method of sequential classification, because it can reduce the cascading errors. Hence, it is challenging to promote the efficiency of online handwriting input, judging from the p@1 evaluation results. From Table XII and Fig. 6, we can observe that DSCIFN still can not handle extremely irregular and sloppy handwritten characters well when the strokes of characters are continuously reduced throughout a sentence. The above analyses indicate that DSCIFN is robust and efficient enough to be used in normal Chinese online handwriting input scenarios, yet the accuracy and robustness still have much room for improvement when poorly handwritten characters are presented.

Second, we select 300 sentences with a length of 30 words, and count the minimum number of strokes at each position that are recognized accurately by models. From the statistical results shown in Fig. 10, we observe that each line fluctuates up and down, which is caused by characters at different positions having different numbers of strokes. Moreover, the increases and decreases of the DSCIFN curve are related to the short-distance dependency of phrases in Chinese sentences. That is, Chinese phrases are composed of multiple Chinese characters, such as, two-character phrases (e.g., ‘陌生’: unfamiliar, ‘真诚’: sincere), three-character phrases (e.g., ‘三角形’: triangles, ‘山水画’: landscape paintings), and four-character phrases (e.g., ‘海阔天空’: as boundless as the sea and sky, ‘万紫千红’: a riot of colour), etc. As shown by one sample in Fig. 11, the purple box indicates that it can still be recognized accurately even if sometimes there are no written strokes. The light blue boxes represent that DSCIFN can accurately recognize Chinese characters written with only one stroke. Many characters only need at

most one stroke to be recognized accurately by DSCIFN, (e.g., ‘格’: lattice, ‘了’: end). Hence, compared with the SSDCNN and VCN (xfmr), DSCIFN can recognize handwritten Chinese characters with fewer strokes precisely in most positions, which promotes the recognition efficiency.

## VI. CONCLUSION

In this paper, we explore how to design precise, robust and effective methods for online handwritten Chinese character recognition by incorporating contextual information. First, to address the lack of a large-scale online handwritten Chinese character dataset with contextual information, we collect a amount of Chinese sentences and hundreds of handwritten character sets to construct a large-scale online handwritten Chinese character with its previous context dataset, named OHCCC, by character matching. To integrate the contextual information of handwritten Chinese characters and their handwriting information, we first propose a vanilla compositional network (VCN) coupled deep convolutional network with a language model, and further design a deep spatial & contextual information fusion network (DSCIFN) in view of the VCN, which heavily depends on contextual information. Finally, we conduct extensive experiments to evaluate the accuracy, robustness and efficiency of the proposed models. DSCIFN is found to perform the best and to achieve excellent performance under all scenarios. Even when the strokes of handwritten Chinese characters continue to decrease with the order of handwriting input, DSCIFN still achieves 97% accuracy. Overall, DSCIFN is a robust and efficient approach for online Chinese handwriting input recognition. For implementation, DSCIFN can be efficiently implemented in actual applications with the rapid development of large-scale cloud servers.

## ACKNOWLEDGMENT

The authors would like to thank the efforts of the editor and reviewers.

## REFERENCES

- [1] R. Plamondon and S. Srihari, “Online and off-line handwriting recognition: A comprehensive survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, Jan. 2000.
- [2] C.-L. Liu and X.-D. Zhou, “Online Japanese character recognition using trajectory-based normalization and direction feature extraction,” in *Proc. 10th Int. Workshop Front. Handwriting Recognit. Suvisoft*, 2006, pp. 1–7.
- [3] R. Plamondon and S. N. Srihari, “Online and off-line handwriting recognition: A comprehensive survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, Jan. 2000.
- [4] S. Lai, L. Jin, and W. Yang, “Toward high-performance online HCCR: A CNN approach with dropdistortion, path signature and spatial stochastic max-pooling,” *Pattern Recognit. Lett.*, vol. 89, pp. 60–66, 2017.
- [5] V. A. Naik and A. A. Desai, “Online handwritten Gujarati character recognition using SVM, MLP, and K-NN,” in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol.*, 2017, pp. 1–6.
- [6] X. Liu, B. Hu, Q. Chen, X. Wu, and J. You, “Stroke sequence-dependent deep convolutional neural network for online handwritten Chinese character recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4637–4648, Nov. 2020.
- [7] D.-H. Wang, C.-L. Liu, and X.-D. Zhou, “An approach for real-time recognition of online chinese handwritten sentences,” *Pattern Recognit.*, vol. 45, no. 10, pp. 3661–3675, 2012.



- [8] S. Quiniou, F. Bouteruche, and E. Anquetil, "Word extraction associated with a confidence index for online handwritten sentence recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 5, pp. 945–966, 2009.
- [9] 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.*, 2011, pp. 37–41.
- [10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [11] K. Jing and J. Xu, "A survey on neural network language models," *CoRR*, vol. abs/1906.03591, 2019.
- [12] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*.
- [13] A. Vaswani *et al.*, "Attention is all you need," in *Advances Neural Inf. Process. Syst.*, vol. 30, Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017.
- [14] S. Zhou, Q. Chen, and X. Wang, "HIT-OR3C: An opening recognition corpus for chinese characters," in *Proc. 9th IAPR Int. Workshop Document Anal. Syst.* New York, NY, USA: Assoc. Comput. Machinery, 2010, pp. 223–230.
- [15] L. Jin, Y. Gao, G. Liu, Y. Li, and K. Ding, "SCUT-COUCH2009-a comprehensive online unconstrained Chinese handwriting database and benchmark evaluation," *Int. J. Document Anal. Recognit.*, vol. 14, pp. 53–64, Mar. 2011.
- [16] C. Liu, S. Jaeger, and M. Nakagawa, "Online recognition of chinese characters: The state-of-the-art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 198–213, Feb. 2004.
- [17] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognit.*, vol. 61, pp. 348–360, 2017.
- [18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
- [19] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, pp. 120–131, 2014.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [22] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.
- [23] A. Byerly, T. Kalganova, and I. Dear, "A branching and merging convolutional network with homogeneous filter capsules," 2020, *arXiv:2001.09136*.
- [24] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 221–233, Jun. 2019.
- [25] Y. Hou, H. Zhang, and S. Zhou, "Convolutional neural network-based image representation for visual loop closure detection," in *Proc. IEEE Int. Conf. Inf. Automat.*, 2015, pp. 2238–2245.
- [26] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
- [27] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3642–3649.
- [28] F. Yin, Q. Wang, X. Zhang, and C. Liu, "ICDAR 2013 chinese handwriting recognition competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, 2013, pp. 1464–1470.
- [29] A. Kawamura *et al.*, "Online recognition of freely handwritten japanese characters using directional feature densities," in *Proc. 11th IAPR Int. Conf. Pattern Recognit.*, vol. II, 1992, pp. 183–186.
- [30] Z.-L. Bai and Q. Huo, "A study on the use of 8-directional features for online handwritten chinese character recognition," in *Proc. 8th Int. Conf. Document Anal. Recognit.*, vol. 1, 2005, pp. 262–266.
- [31] X. Qu, N. Xu, W. Wang, and K. Lu, "Similar handwritten Chinese character recognition based on adaptive discriminative locality alignment," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl.*, 2015, pp. 130–133.
- [32] T. V. Phan, J. Gao, B. Zhu, and M. Nakagawa, "Effects of line densities on nonlinear normalization for online handwritten Japanese character recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, 2011, pp. 834–838.
- [33] Z. Xu, X. Wang, and S. Jiang, "A sentence-level Chinese character input method," *High Technol. Lett.*, vol. 1, 2000.
- [34] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio, "Drawing and recognizing Chinese characters with recurrent neural network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 849–862, Apr. 2018.
- [35] A. Das, G. R. Patra, and M. N. Mohanty, "LSTM based Odia handwritten numeral recognition," in *Proc. Int. Conf. Commun. Signal Process.*, 2020, pp. 0538–0541.
- [36] H. Q. Ung, C. T. Nguyen, H. T. Nguyen, T.-N. Truong, and M. Nakagawa, "A transformer-based math language model for handwritten math expression recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, 2021, pp. 403–415.
- [37] V. Carbune *et al.*, "Fast multi-language lstm-based online handwriting recognition," *Int. J. Document Anal. Recognit.*, vol. 23, no. 2, pp. 89–102, 2020.
- [38] S. Gehrmann, Y. Deng, and A. Rush, "Bottom-up abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Oct./Nov. 2018, pp. 4098–4109.
- [39] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/Daily mail reading comprehension task," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, Aug. 2016, pp. 2358–2367.
- [40] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789.
- [41] B. N. Patro, V. K. Kurmi, S. Kumar, and V. Nambodiri, "Learning semantic sentence embeddings using sequential pair-wise discriminator," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2715–2729.
- [42] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *CoRR*, vol. abs/1703.09902, 2017.
- [43] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [44] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.
- [45] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [46] M. Lewis *et al.*, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 7871–7880.
- [47] C. Yan *et al.*, "Stat: Spatial-temporal attention mechanism for video captioning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 229–241, Jan. 2020.
- [48] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.
- [49] Y. Meng *et al.*, "Glyce: Glyph-vectors for Chinese character representations," in *Advances Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019.
- [50] J. Williams, S. Kleinogesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal DNN feature fusion," in *Proc. Grand Challenge Workshop Human Multimodal Lang. (Challenge-HML)*. Melbourne, Australia: Assoc. Comput. Linguist., Jul. 2018, pp. 11–19.
- [51] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Sep. 2017, pp. 1103–1114.
- [52] A. Zadeh *et al.*, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
- [53] Z. Liu *et al.*, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, vol. 1. Melbourne, Australia: Assoc. Comput. Linguist., Jul. 2018, pp. 2247–2256.
- [54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [55] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Conf. Neural Inf. Process. Syst. Workshop Deep Learn.*, Dec. 2014.
- [56] A. F. Agarap, "Deep learning using rectified linear units (relu)," 2018, *arXiv:1803.08375*.
- [57] T. Iqbal and S. Qureshi, "The survey: Text generation models in deep learning," *J. King Saud Univ. - Comput. Inf. Sci.*, pp. 1319–1578, 2020.
- [58] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ, USA: Prentice Hall, 1994.
- [59] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *CoRR*, vol. abs/1811.03378, 2018.

- [60] W. Yang, L. Jin, D. Tao, Z. Xie, and Z. Feng, "Dropsample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition," *Pattern Recognit.*, vol. 58, pp. 190–203, 2016.
- [61] X.-Y. Zhang and C.-L. Liu, "Writer adaptation with style transfer mapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1773–1787, Jul. 2013.



**Baotian Hu** received the M.S. and Ph.D. degrees in computer science from the Shenzhen Graduate School, Harbin Institute of Technology (Shenzhen), Shenzhen, China, in 2012 and 2016, respectively. He is currently an Assistant Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). His current research interests include deep learning and its application on natural language processing and image recognition.



**Yunxin Li** received the B.S. degree from the School of Information and Computing Science, Harbin Institute of Technology, Weihai, China, in 2019. He is currently working toward the M.S. degree with the School of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include vision and language, cross-modal, and natural language generation.



**Xiaolong Wang** received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1989. He is currently a Full Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests include deep learning and its application on natural language processing and sentence-level Pinyin input method.



**Qian Yang** received the B.E. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2020. She is currently working toward the M.S. degree with the School of Computer Science, Harbin Institute of Technology (Shenzhen), Shenzhen, China. Her research interests include vision and language, and generative models.



**Yuxin Ding** received the Ph.D. degree in computer software from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His current research interests include deep learning and natural language processing.



**Qingcai Chen** (Member, IEEE) received the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1998 and 2003, respectively. He is currently a Professor and the Director of the Center for Intelligent Computing Research, School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include natural language processing, artificial intelligence, machine learning, financial, and medical information processing.



**Lin Ma** received the B. E., and M. E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, in 2013. He is currently a Researcher with Meituan, Beijing, China. His current research interests include deep learning and multimodal learning, specifically for image and language.