This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

Recognition of Handwritten Chinese Text by Segmentation: A Segment-annotation-free Approach

Dezhi Peng, Lianwen Jin, Weihong Ma, Canyu Xie, Hesuo Zhang, Shenggao Zhu, and Jing Li

***Abstract*—Online and offline handwritten Chinese text recognition (HTCR) has been studied for decades. Early methods adopted oversegmentation-based strategies but suffered from low speed, insufficient accuracy, and high cost of character segmentation annotations. Recently, segmentation-free methods based on connectionist temporal classification (CTC) and attention mechanism, have dominated the field of HCTR. However, people actually read text character by character, especially for ideograms such as Chinese. This raises the question: are segmentation-free strategies really the best solution to HCTR? To explore this issue, we propose a new segmentation-based method for recognizing handwritten Chinese text that is implemented using a simple yet efficient fully convolutional network. A novel weakly supervised learning method is proposed to enable the network to be trained using only transcript annotations; thus, the expensive character segmentation annotations required by previous segmentation-based methods can be avoided. Owing to the lack of context modeling in fully convolutional networks, we propose a contextual regularization method to integrate contextual information into the network during the training stage, which can further improve the recognition performance. Extensive experiments conducted on four widely used benchmarks, namely CASIA-HWDB, CASIA-OLHWDB, ICDAR2013, and SCUT-HCCDoc, show that our method significantly surpasses existing methods on both online and offline HCTR, and exhibits a considerably higher inference speed than CTC/attention-based approaches.**

***Index Terms*—Handwritten Chinese text recognition, Online and offline text recognition, Segmentation-based text recognition, Weakly supervised learning**

I. Introduction

HANDWRITTEN Chinese text recognition (HCTR), including online and offline HCTR, is a challenging research topic that has been intensively studied for decades. However, owing to the large vocabulary (tens of thousands of character categories), diverse writing styles, and the character-touching problem, satisfactory recognition performance has not yet been achieved.

Early methods of HCTR are based on oversegmentation [1], [2], [3], [4], [5], [6], [7], [8], which first oversegment the text line and then search for the best segmentation-recognition path by integrating classifier outputs, geometric context, and linguistic context. Oversegmentation-based methods used to be the most successful approaches for HCTR; however, they can be easily affected by touching or overlapping characters and require annotations to provide the boundary of characters. Recently, segmentation-free methods [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24] based on hidden Markov model (HMM), connectionist temporal classification (CTC) [25], and attention mechanism [26], have dominated the field of HCTR. They surpass

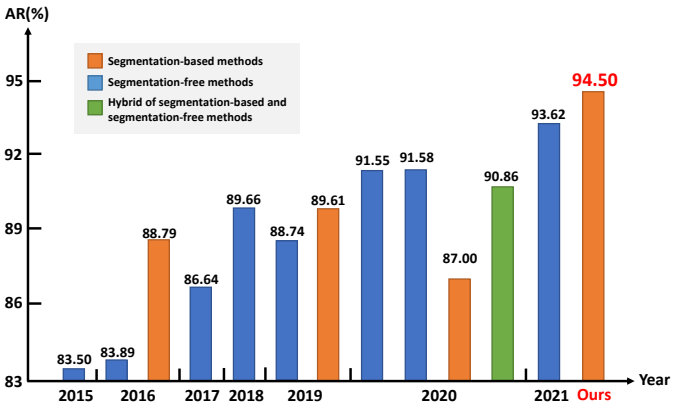
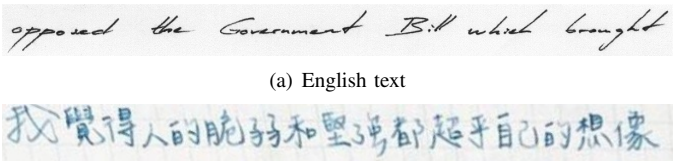


Fig. 1. Performances of typical methods on the offline subset of ICDAR2013 competition dataset. Our method re-explores the segmentation-based pipeline and makes a significant improvement.

oversegmentation-based methods and only require transcripts to be annotated. We further illustrate the performance of typical methods on the widely used offline subset of the ICDAR2013 competition dataset [4] in Fig. 1. Although the segmentation-based method achieved considerable improvement before 2016, segmentation-free methods have become mainstream. However, are segmentation-free approaches really the best solutions to the HCTR problem? Fig. 2 shows two example images of handwritten English text from the IAM dataset [27] and handwritten Chinese text from the SCUT-HCCDoc dataset [28]. In contrast to English texts, the basic elements of Chinese texts are characters rather than words. Intuitively, for a native Chinese speaker, each character is first segmented from the text line and then recognized when reading Chinese texts. In addition, compared with English characters, Chinese characters have a large vocabulary, diverse writing styles, complicated two-dimensional structures, and serious imbalance of character frequency, which can easily cause misalignment in segmentation-free methods, especially attention-based approaches [29], [30]. Furthermore, character segmentation information can help many downstream tasks, such as text removal, text editing, and visual information extraction. However, segmentation-free methods cannot explicitly produce character segmentation results, which limits their potential for many applications.

In this paper, following our previous work [31], we explore a new method for recognizing handwritten Chinese text by segmentation. *First, we formulate a new segmentation-based text recognition framework for HCTR which end-to-end segments and recognizes characters through a fully convo-*

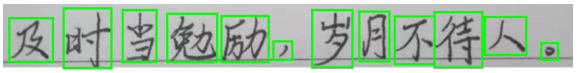
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia



(a) English text

(b) Chinese text

Fig. 2. Example images from IAM (English) and SCUT-HCCDoc (Chinese) datasets.



(a) Character segmentation (52s)

及 时 当 勉 励 ， 岁 月 不 待 人 。

(b) Transcript (5s)

Fig. 3. Different annotations and their time cost measured by the LabelMe[1] tool.

lutional network. Compared with previous oversegmentation-based methods, the proposed framework is end-to-end trainable with high efficiency. Moreover, compared with most existing segmentation-free methods that utilize recurrent neural networks (RNNs) or auto-regressive models, our method runs in parallel and exhibits a higher inference speed. *Second, we propose a weakly supervised learning method to enable the network to be trained using only transcript annotations.* Most existing segmentation-based methods [1], [2], [3], [4], [5], [6], [7], [31] require expensive character segmentation annotations, i.e., character bounding boxes, which are evidently much more tedious and time-consuming than transcript annotations. However, our method avoids such costly annotations and can still output character segmentation results. The illustration and time cost of different annotations are shown in Fig. 3. *Third, a new contextual regularization method is proposed to integrate contextual information by guiding the feature extraction of the network.* The performance can be further boosted through the proposed contextual regularization.

Compared with the conference version [31], the major extension of this paper lies in the weakly supervised method and contextual regularization, which significantly reduce the cost of manual annotation and improve the recognition performance, respectively. Moreover, the conference version [31] only focused on offline HCTR, whereas we further verified the effectiveness of our method on online HCTR and camera-captured handwritten text recognition.

The experiments are conducted using CASIA-HWDB [33], CASIA-OLHWDB [33], ICDAR2013 [4], and SCUT-HCCDoc [28]. Our method achieves state-of-the-art performance on these datasets, which demonstrates the success of our method on both online and offline HCTR. Moreover, additional experiments conducted using ReCTS-25k [34] demonstrate the potential of our approach for scene text recognition. We hope that this paper can inspire more research to re-explore segmentation-based methods in addition to CTC and attention mechanism.

To summarize, the main contributions of this paper are as follows:

- We formulate a new segmentation-based text recognition framework for online and offline HCTR. During inference, the framework is implemented using a fully convolutional network, and thus exhibits high efficiency.
- We propose a weakly supervised learning method to enable the network to be trained using only transcript annotations, thereby greatly reducing the cost of manual annotations. However, character segmentation results can still be produced using our method.
- We design a contextual regularization method to integrate contextual information into the training of the fully convolutional network without slowing down the inference speed.
- Extensive experiments demonstrate the state-of-the-art performance of our method on both online and offline HCTR, in terms of both recognition accuracy and inference speed.

II. Related Work

A. Offline Handwritten Chinese Text Recognition

Offline HCTR aims to transcribe text-line images into Chinese texts. In general, existing methods can be categorized into segmentation-based and segmentation-free approaches.

Most previous segmentation-based (also called explicit-segmentation) approaches adopt oversegmentation-based strategies. Specifically, these methods [1], [6] first obtain primitive segments through oversegmentation and then search for the best segmentation-recognition path. Although the character classifier, language model, and geometric model are integrated for path evaluation, satisfactory performance cannot be achieved, especially for touching or overlapping characters. Therefore, Wu et al. [7] explored neural network language models, yielding a significant improvement in the recognition performance. Furthermore, Wang et al. [8] proposed a weakly supervised method for string-level training of oversegmentation-based systems. In addition to the strategy using oversegmentation, Peng et al. [31] designed a fully convolutional network for end-to-end handwritten Chinese text segmentation and recognition.

Although enormous progress has been achieved by previous segmentation-based methods, most of them require expensive character segmentation annotations. To address this issue, segmentation-free (also called implicit-segmentation) approaches have recently received considerable interest. One category of segmentation-free methods [9], [10], [11], [12] addresses the offline HCTR problem using hidden Markov model (HMM), where the handwritten text lines are modeled by a series of cascading HMMs based on the feature sequence extracted in a sliding-window manner. The method based on CTC [25] is another popular category of solutions to offline HCTR. Messina et al. [15] solved the offline HCTR problem by combining multi-dimensional long short-term memory (MDLSTM) and CTC. Wu et al. [17] further proposed separable MDLSTM to improve efficiency. Xie et al. [21]

[1]https://github.com/wkentaro/labelme

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

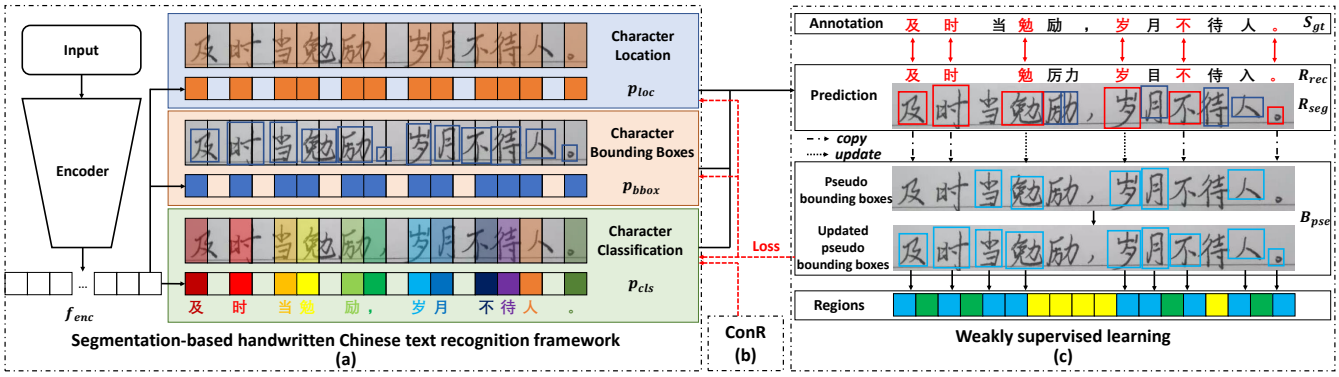


Fig. 4. Proposed method consists of three parts: (a) a new segmentation-based handwritten Chinese text recognition framework, (b) a contextual regularization (ConR) term for integrating contextual information, and (c) a weakly supervised learning method for training the model using only transcript annotations. The red dashed lines indicate the optimization process at training stage.

improved the CTC-based methods by exploring data augmentation and preprocessing. Liu et al. [65] explored residual and squeeze-and-excitation structures for feature extraction and proposed context beam search to integrate the Transformer-based [63] language model into CTC-based methods. The attention mechanism, which has been widely adopted in scene text recognition [35], [30], [67], action recognition [37], [38], and video processing [39], can also be applied to offline HCTR. Xiu et al. [40] improved the attention-based decoder by a multi-level multi-modal fusion network. In addition to the above methods that adopt a single strategy, Zhu et al. [22] proposed a convolutional combination strategy to combine the segmentation-based and segmentation-free approaches for better performance.

B. Online Handwritten Chinese Text Recognition

Online HCTR is aimed at recognizing Chinese text from pen-tip trajectories. In contrast to offline HCTR, the input of online HCTR contains sequential information and does not have background noise. However, the acquisition of pen-tip trajectories requires specific hardware. Owing to the rise of pen-based devices, online HCTR is also a very important research topic with wide applications in many fields such as finance and education.

The oversegmentation-based strategy can also be used to address online HCTR [2], [3], [5], following the same pipeline as for offline HCTR. However, these methods rely heavily on the evaluation of candidate segmentation-recognition paths and have difficulty in recognizing touching or overlapping characters. Therefore, some studies [13], [41], [19] proposed to directly handle the pen-tip trajectory using RNNs and CTC. These methods extract features using long short-term memory (LSTM) or gated recurrent unit (GRU) and optimize the model with CTC loss. Liu et al. [19] further proposed distilling GRU to accelerate model training and handle handwritten texts with various styles. Instead of using RNNs, Peng et al. [64] designed a global and local relation network (GLRNet) that uses self-attention [63] for feature extraction and jointly trained the combination of GLRNet and a Transformer-based language model to achieve optimal overall performance. With the prevalence of convolutional neural networks (CNNs), there exist methods [32], [16], [18] that transform the online pen-tip trajectory into offline feature maps that CNNs can process. Such methods first obtain image-like representations through path signature or eight-directional feature maps. Thereafter, an integrated CNN-LSTM network is trained using CTC loss. Xie et al. [32] further improved this pipeline using an implicit language model and multi-spatial context.

III. Methodology

A. Overview

The overall design of the proposed method is illustrated in Fig. 4. First, a new segmentation-based handwritten Chinese text recognition framework that segments and recognizes characters is formulated in an end-to-end manner and is implemented using an efficient fully convolutional network during inference. Second, a novel weakly supervised learning method is proposed to effectively train the model using only transcript annotations, which avoids costly manual segmentation annotations. Nevertheless, our model can still accurately output character segmentation results. Third, as the fully convolutional architecture can not capture contextual dependencies, a new contextual regularization (ConR) is proposed to integrate contextual information into the model during the training stage. The proposed ConR can improve the performance without reducing the inference speed.

B. Segmentation-based Text Recognition

The proposed segmentation-based text recognition framework is illustrated in Fig. 4(a). The input is text-line images for offline HCTR or image-like representations generated from pen-tip trajectories for online HCTR. For simplicity and clarity, we take offline HCTR as an example in Fig. 4

Given an input $I_{in} \in \mathbb{R}^{H\times W\times C}$ (H, W, and C denote the height, width, and number of channels of the input, respectively), the encoder extracts the feature map f_{enc} as

$$f_{enc} \in \mathbb{R}^{1\times w_{enc}\times c_{enc}} = Encoder(I_{in}), \tag{1}$$

where c_{enc} and w_{enc} are the number of channels and the width of the feature map, respectively. The height of the feature map f_{enc} is downsampled to one, and the $Encoder$ is a CNN.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

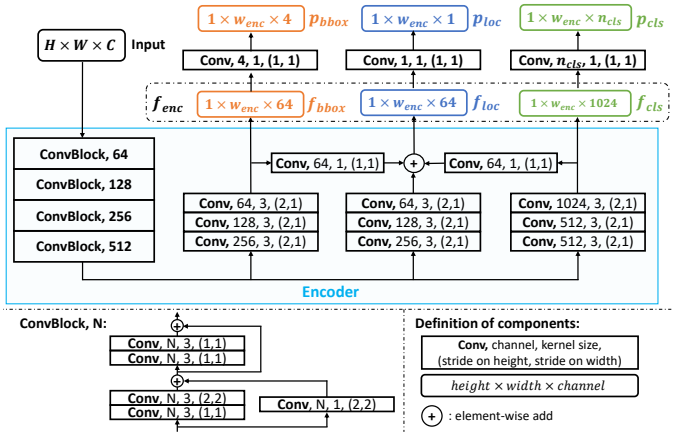


Fig. 5. Detailed architecture of our network.

Then, three predictions are produced based on the feature map f_{enc} as

$$p_{loc} \in \mathbb{R}^{w_{enc} \times 1}, p_{bbox} \in \mathbb{R}^{w_{enc} \times 4}, p_{cls} \in \mathbb{R}^{w_{enc} \times n_{cls}}, \quad (2)$$

where p_{loc}, p_{bbox}, and p_{cls} are the character location, character bounding boxes, and character classification, respectively. The n_{cls} is the total number of character categories. Based on these predictions, input I_{in} is equally divided into w_{enc} regions. The p^n_{loc} is the confidence that the n-th region contains characters, p^n_{bbox} is the coordinates of the bounding box of the character in the n-th region, and p^n_{cls} contains the probabilities that the character in the n-th region is classified as each of the n_{cls} categories.

We implement this framework using a fully convolutional network during inference. The detailed network architecture is illustrated in Fig. 5. Inspired by ResNet [42], the residual connection is adopted in the *ConvBlock*. Instead of extracting a single feature map f_{enc}, the encoder outputs three feature maps, f_{loc}, f_{bbox}, and f_{cls}, to predict the character location, character bounding boxes, and character classification, respectively.

During inference, after removing redundant predictions using non-maximum suppression (NMS) [43] and sorting the remaining predictions from left to right, we can obtain segmentation and recognition results as shown in the "Prediction" part of Fig. 4(c). Specifically, the score of the character in the n-th region during NMS is the weighted sum of p^n_{loc} and the maximum probability in p^n_{cls}, in order to integrate semantic information into the character location. As specified in the conference version [31], the weight of p^n_{loc} is set to 0.8.

C. Weakly Supervised Learning

In this section, we propose a weakly supervised learning method (Fig. 4(c)) to enable our network to be trained using only transcript annotations; nevertheless, our method can still output character segmentation results that cannot be produced by segmentation-free methods.

Although character segmentation annotations are costly to annotate, it is easy to obtain isolated Chinese character samples from font files. Using these cost-free isolated character samples, we can easily synthesize text lines with character segmentation annotations. However, simply training the network using synthetic data cannot achieve satisfactory performance on real text lines, especially in complex scenarios. One solution is to develop advanced synthesis methods [44], [45], [46], [47] to better mitigate the real text lines. However, for different scenarios, different data should be synthesized, and even different methods should be designed. Moreover, most existing synthesis methods produce only transcript annotations. Therefore, in our weakly supervised learning method, we proposed to exploit useful information from simple synthetic data to help the model learn from real data under the guidance of transcripts.

Specifically, the network is first pretrained using synthetic data with segmentation annotations. As described in Section IV-B, the samples for different scenarios are synthesized in the same manner, by placing isolated characters on a white background without any complicated synthesis techniques. Although the synthetic data may differ significantly from the text lines from real scenes, the model can still learn the fundamental ability to localize and recognize characters.

Then, the real data with only transcript annotations is also used to train the model. The procedure of the proposed weakly supervised learning method for real samples is shown in Fig. 4(c). Similar to the learning process of humans, the model is taught which prediction is correct and trained based on past successful experiences. Specifically, for a real input, the network predicts the segmentation result R_{seg} and recognition result R_{rec} as

$$R_{seg} = \{(r^1_{seg}, r^1_{sco}), (r^2_{seg}, r^2_{sco}), ..., (r^{l_{pr}}_{seg}, r^{l_{pr}}_{sco})\}, \quad (3)$$

$$R_{rec} = \{r^1_{rec}, r^2_{rec}, ..., r^{l_{pr}}_{rec}\}, \quad (4)$$

where r^i_{seg}, r^i_{sco}, and r^i_{rec} are the coordinates of the bounding box, score, and category of the i-th predicted character, respectively, and l_{pr} is the total number of predicted characters.

Thereafter, we compute the edit distance between transcript annotation S_{gt} and recognition result R_{rec}, where S_{gt} is defined as

$$S_{gt} = \{s^1_{gt}, s^2_{gt}, ..., s^{l_{gt}}_{gt}\}, \quad (5)$$

where s^j_{gt} is the category of the j-th character in the transcript, and l_{gt} is the total number of annotated characters. The characters, which are matched as "equal" in computing edit distance, are marked using red arrows between the annotation and prediction in Fig. 4(c). Because character recognition and segmentation tasks are highly related, the "equal" characters in the prediction are very likely to have accurate bounding box predictions.

Next, the bounding boxes corresponding to "equal" characters (red bounding boxes in Fig. 4(c)) are used to update the pseudo bounding boxes. The pseudo bounding boxes B_{pse} are defined as follows:

$$B_{pse} = \{(b^1_{pse}, b^1_{sco}), (b^2_{pse}, b^2_{sco}), ..., (b^{l_{gt}}_{pse}, b^{l_{gt}}_{sco})\}, \quad (6)$$

where b^j_{pse} and b^j_{sco} are the coordinates and score of the pseudo bounding box of character s^j_{gt}, respectively. All the pseudo bounding boxes are initialized as $\varnothing$. If the character r^i_{rec} is matched to be "equal" to the character s^j_{gt}, the segmentation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

result r^i_{seg} is used to update the pseudo bounding box as follows:

$$b^j_{pse} = \begin{cases} r^i_{seg}, & b^j_{pse} = \varnothing, \\ \lambda_{pse} * b^j_{pse} + (1-\lambda_{pse}) * r^i_{seg}, & otherwise, \end{cases} \tag{7}$$

$$b^j_{sco} = \begin{cases} r^i_{sco}, & b^j_{pse} = \varnothing, \\ \lambda_{pse} * b^j_{sco} + (1-\lambda_{pse}) * r^i_{sco}, & otherwise, \end{cases} \tag{8}$$

where λ_{pse} is calculated as:

$$\lambda_{pse} = \frac{e^{10\times b^j_{sco}}}{e^{10\times b^j_{sco}} + e^{10\times r^i_{sco}}}. \tag{9}$$

The weight λ_{pse} is a function of b^j_{sco} and r^i_{sco}, so as to make the bounding box with a higher score have a much higher weight.

Subsequently, the pseudo bounding boxes B_{pse} and transcript annotation S_{gt} are used to optimize the network. However, there may exist pseudo bounding boxes equal to $\varnothing$, which implies that the loss cannot be computed in a normal manner. Therefore, a special method for loss calculation is designed.

Specifically, we first project the pseudo bounding boxes that are not equal to $\varnothing$ onto their corresponding regions of the model input, yielding a mapping M_{ptr}. The element $(j,n) \in M_{ptr}$ indicates that the center point of the pseudo bounding box b^j_{pse} is within the n-th region. The regions that contain the center points of pseudo bounding boxes are represented by the blue squares in the "Regions" part of Fig. 4(c). Then the loss of p_{bbox} and p_{cls} are calculated as:

$$l_{bbox} = \frac{1}{|M_{ptr}|} \sum_{(j,n)\in M_{ptr}} SE(b^j_{pse}, p^n_{bbox}), \tag{10}$$

$$l_{cls} = -\frac{1}{|M_{ptr}|} \sum_{(j,n)\in M_{ptr}} log(p^{n,s^j_{gt}}_{cls}), \tag{11}$$

where the function SE calculates the square error between two inputs.

Regarding the loss of character location p_{loc}, we only know that the blue regions in Fig. 4(c) contain characters. Owing to the existence of pseudo bounding boxes equal to $\varnothing$, the difficulty lies in determining the negative samples, i.e., the regions that do not contain characters. Fortunately, although we cannot find all the negative samples, it can be confirmed that there is no character in the regions between two consecutive pseudo bounding boxes. Specifically, if both b^j_{pse} and b^{j+1}_{pse} are not equal to $\varnothing$, the indices of the regions, which are between the two regions corresponding to b^j_{pse} and b^{j+1}_{pse}, are added to the set N_{loc}. The regions in N_{loc} contain no characters and are represented by the green squares in Fig. 4(c). However, whether the yellow regions contain characters cannot be determined because of missing pseudo bounding boxes. Thus, these regions are not considered in the loss calculation. Consequently, the loss of p_{loc} is formulated as:

$$l_{loc} = -\frac{0.5}{|T_{loc}|} \sum_{n\in T_{loc}} log(p^n_{loc}) - \frac{0.5}{|N_{loc}|} \sum_{n\in N_{loc}} log(1-p^n_{loc}), \tag{12}$$

where T_{loc} is the set of the indices of the blue regions that are supposed to contain characters.

Finally, the total loss is given by:

$$l_{total} = l_{bbox} + l_{cls} + l_{loc}. \tag{13}$$

D. Contextual Regularization

The fully convolutional architecture results in high efficiency and parallel computing, but it is difficult to capture contextual dependencies. The prediction for each region of a text line depends only on the corresponding receptive field. For example, the character classification p_{cls} can be reformulated as:

$$\begin{aligned} p_{cls} &= \{p^1_{cls}, ..., p^{w_{enc}}_{cls}\} = FCN(I_{in}) \\ &= \{FCN(F^1), ..., FCN(F^{w_{enc}})\}, \end{aligned} \tag{14}$$

where FCN is our network and F^n is the receptive field of the n-th character classification prediction. In the above formula, the predictions in p_{cls} are independent without considering the relationship between them.

However, the task of HCTR, especially the character classification part, relies heavily on contextual information. Even for native speakers, it is difficult to recognize some confusing cursive handwritten characters without context. Previous methods adopted recurrent layers, such as LSTM and GRU, for context modeling. However, these recurrent layers can not run in parallel, which greatly reduces the inference speed especially when handling long texts. To this end, we propose to solve context modeling from a new perspective, by guiding the feature extraction using a novel contextual regularization (ConR) only in the training stage, as illustrated in Fig. 6.

During training, two bidirectional long short-term memory (BLSTM) layers are added on top of the feature map f_{cls}. Subsequently, an extra character classification is predicted based on the output of the BLSTM layers. The loss l_{conr} of this new character classification is calculated in the same way as for l_{cls} (Eq. (11)). Then, the loss l_{conr} is added to the total loss l_{total} as a regularization term. Through the backpropagation of the gradient, ConR can guide the feature f_{cls} to capture contextual information with the help of the context modeling ability of BLSTM layers.

During inference, the BLSTM layers and subsequent character classification are removed to maintain the high processing speed of the fully convolutional architecture. Experiments show that ConR results in a consistent improvement on all datasets. Even if the BLSTM layers and subsequent character classification are used during inference, the performance remains nearly unchanged, which proves that the feature map f_{cls} really learns contextual information.

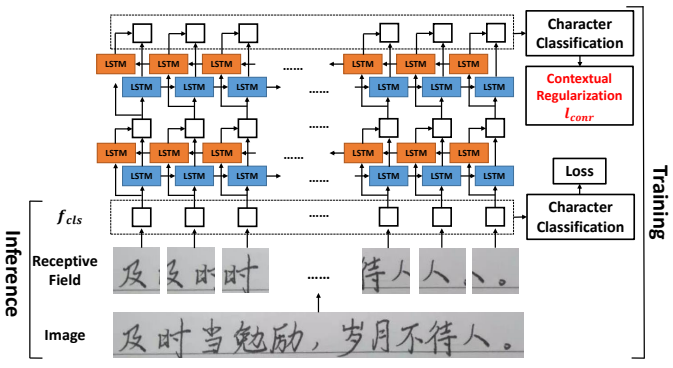
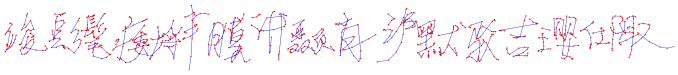
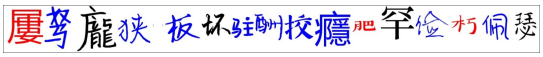


Fig. 6. Illustration of the contextual regularization.

(a) For CASIA-HWDB



(b) For CASIA-OLHWDB



(c) For SCUT-HCCDoc

Fig. 7. Examples of synthetic samples (without semantic information) for CASIA-HWDB, CASIA-OLHWDB, and SCUT-HCCDoc.

IV. Experiments

A. Datasets

CASIA-HWDB [33] is a large-scale offline handwriting database, including CASIA-HWDB1.0-1.2 and CASIA-HWDB2.0-2.2. CASIA-HWDB1.0-1.2 contain 3,895,135 isolated characters from 1,020 writers, while CASIA-HWDB2.0-2.2 include 52,230 text lines from 1,019 writers. Note that the isolated character samples of CASIA-HWDB1.0-1.2 are not the characters cropped from the text lines of CASIA-HWDB2.0-2.2.

CASIA-OLHWDB [33] is the online version of CASIA-HWDB, which consists of CASIA-OLHWDB1.0-1.2 and CASIA-OLHWDB2.0-2.2. CASIA-OLHWDB1.0-1.2 contain 3,912,017 isolated characters from 1,020 writers while CASIA-OLHWDB2.0-2.2 contain 52,220 text lines from 1,019 writers. The CASIA-OLHWDB2.0-2.2 are further divided into 41,710 text lines for training and 10,510 text lines for testing.

ICDAR2013 competition dataset [4] contains 3,432 online and offline handwritten Chinese text lines from 60 writers. For convenience, the online and offline subsets are denoted as **ICDAR2013-Online** and **ICDAR2013-Offline**, respectively.

SCUT-HCCDoc [28] contains 12,253 offline camera-captured document images with 116,629 text lines. The training and testing sets comprise 93,411 text lines and 23,218 text lines, respectively.

B. Data Synthesis

For all real datasets, synthetic text lines are synthesized by simply placing offline characters on a white background or concatenating the pen-tip trajectories of online characters. The synthetic samples for CASIA-HWDB and CASIA-OLHWDB use isolated characters from CASIA-HWDB1.0-1.2 and CASIA-OLHWDB1.0-1.2, respectively, while the synthetic samples for SCUT-HCCDoc use the character images generated from 101 font files. The categories of characters are randomly selected from the vocabulary when synthesizing data without semantic information, while the corpora described in Section IV-D are adopted when synthesizing data with semantic information. Fig. 7 shows examples of synthetic samples (without semantic information) for CASIA-HWDB, CASIA-OLHWDB, and SCUT-HCCDoc.

C. Implementation Details

We conduct our experiments using an NVIDIA GTX 1080ti GPU with 11GB of memory and implement our method using PyTorch. First, in the pretraining stage, the model is pretrained using synthetic data for 150,000 iterations. The weakly supervised learning method and ConR are not adopted in this stage. Then, in the training stage, the model is trained using both real and synthetic data for 1,200,000 iterations. The synthetic data is synthesized on the fly. The network is optimized using stochastic gradient descent (SGD) with a batch size of 8 and an initial learning rate of 0.01. The learning rate is multiplied by 0.1 at 25%, 50%, and 75% of the total number of iterations. During inference, the batch size is set to 1.

D. Transcription

During inference, the transcription without language model is described in the last paragraph of Section III-B, which first removes redundant predictions through NMS and then rearranges the remaining characters from left to right. In the following sections, if not specified, the results are obtained by the transcription without language model.

Moreover, the transcription process can be combined with n-gram language models. Specifically, a tri-gram language model generated from the same corpora as the conference version [31] is adopted. The corpora consist of the PFR corpus [48] (news text of 2,199,492 characters from the 1998 People's Daily corpus), the PH corpus [49] (news text of 3,697,028 characters from the People's Republic of China's Xinhua news recorded between January 1990 and March 1991), and the CLDC corpus [50] (50 million characters collected by the Institute of Applied Linguistics). Because the character classification prediction p_{cls} contains the probability of each character category and the blank probability can be calculated as $1 - p_{loc}$, the CTC beam search algorithm [51] can be used for the transcription with the tri-gram language model based on the CTC-style predictions formed by p_{cls} and $1 - p_{loc}$.

Recently, the Transformer-based [63] language model has emerged in the field of HCTR [40], [64], [65]. After preparing CTC-style predictions as mentioned above, the transcription process can also be integrated with the Transformer-based language model through the context beam search algorithm [65]. The Transformer-based language model follows the architecture specified in [65] and is trained with the same corpora as the tri-gram language model using the Fairseq [66] toolkit.

E. Evaluation Metrics

Following previous studies on online and offline HCTR, the accurate rate (AR) and correct rate (CR) are adopted to evaluate the performance of methods, which are calculated as

$$\begin{aligned} AR &= (N_t - D_e - S_e - I_e)/N_t, \\ CR &= (N_t - D_e - S_e)/N_t, \end{aligned} \tag{15}$$

where D_e, S_e, and I_e represent the total number of deletion, substitution, and insertion errors, respectively, and N_t is the total number of characters in the annotations. The errors

are calculated between the recognition result and transcript annotation. The AR and CR are presented as percentages in the tables of the following sections.

F. Experiments on ICDAR2013-Offline

1) Experimental Settings: The model is first pretrained using the synthetic data (without semantic information) for CASIA-HWDB. Then, the 52,230 real samples from CASIA-HWDB2.0-2.2 and the synthetic data (with semantic information) for CASIA-HWDB are used to train the model. Both the ratios of the real and synthetic samples in a batch are 0.5. Finally, the model is evaluated on the 3,432 samples from ICDAR2013-Offline. Following the setting of most previous methods, the number of character categories is set to 7,356.

2) Data Preprocessing: For the text-line images from CASIA-HWDB2.0-2.2 and ICDAR2013-Offline, the data preprocessing is illustrated in Fig. 8. Given a text-line image (Fig. 8(a)), we first estimate the tilt angle of the text by conducting linear regression using the coordinates of black pixels (Fig. 8(b)). Then, the text-line image is rotated to make the text horizontal (Fig. 8(c)). Next, we remove the white padding at the top and bottom of the image to highlight the text (Fig. 8(d)). Finally, the height of the text-line image is normalized to 128 pixels while maintaining the aspect ratio (Fig. 8(e)).

For the synthetic text-line images, only size normalization (Fig. 8(e)) is performed.

3) Experimental Results: Table I shows the comparison of existing methods and ours on ICDAR2013-Offline. It can be seen that our approach achieves state-of-the-art performance with and without language model. Specifically, taking advantage of the Transformer-based language model, the method in [65] outperforms our approach that uses the traditional trigram language model. However, state-of-the-art performance can still be achieved when our method is also equipped with the Transformer-based language model.

Moreover, in addition to the outstanding recognition performance, our method can also output character segmentation results. Fig. 9 shows the visualization results of ICDAR2013-Offline. It can be seen that the characters can be segmented accurately from the text-line image. Particularly in the fourth example, the character "请" is written very similar to "清" but is still correctly recognized (marked by blue square), demonstrating the ability to distinguish similar characters. This ability may be attributed to the contextual information integrated by ConR and the large-scale synthetic datasets. Nonetheless,

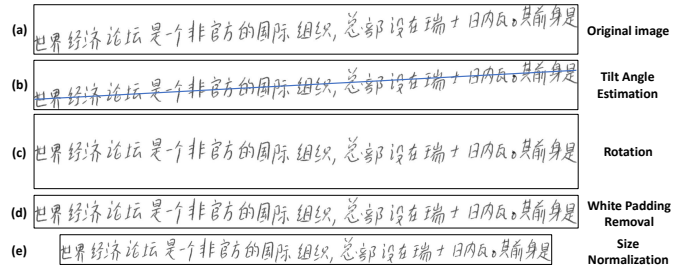


Fig. 8. Illustration of data preproccessing for the text-line images from CASIA-HWDB2.0-2.2 and ICDAR2013-Offline.

TABLE I
COMPARISON WITH EXISTING METHODS ON ICDAR2013-OFFLINE (LM: LANGUAGE MODEL)

Method	Without LM		With LM	
	AR	CR	AR	CR
HIT-2 [4]	-	-	86.73	88.76
Messina et al. [15]	83.50	-	89.40	-
Wu et al. [17]	86.64	87.43	90.38	-
Du et al. [10]	83.89	-	93.50	-
Wang et al. [6]	88.79	90.67	94.02	95.53
Wu et al. [7]	-	-	96.20	96.32
Wang et al. [11]	89.66	-	96.47	-
Xie et al. [20]	91.25	91.68	96.22	96.70
Peng et al. [31]	89.61	90.52	94.88	95.51
Xiu et al. [40]	88.74	-	96.35	-
Xie et al. [21]	91.55	92.13	96.72	96.99
Wang et al. [12]	91.58	-	96.83	-
Wang et al. [8]	87.00	89.12	95.11	95.73
Zhu et al. [22]	90.86	-	94.00	-
Liu et al. [65]	93.62	-	97.51	-
Ours (tri-gram LM)	**94.50**	**94.76**	96.79	97.32
Ours (Transformer-based LM)	**94.50**	**94.76**	**97.70**	**97.91**

there are still some misrecognized similar characters, e.g., "经" is recognized as "往" in the second example, indicating the recognition of similar handwritten Chinese characters is still an important issue worth studying in the future.

G. Experiments on ICDAR2013-Online

1) Experimental Settings: First, the synthetic data (without semantic information) for CASIA-OLHWDB is used to pretrain the model. Then, the model is trained using the 41,710 real samples from the training set of CASIA-OLHWDB2.0-2.2 and the synthetic data (with semantic information) for CASIA-OLHWDB. Both the ratios of the real and synthetic samples in a batch are 0.5. After training, the 3,432 samples from ICDAR2013-Online are used for evaluation. Following previous studies, the number of character categories is set to 7,356.

2) Data Preprocessing: The pipeline of data preprocessing for pen-tip trajectories from CASIA-OLHWDB2.0-2.2 and

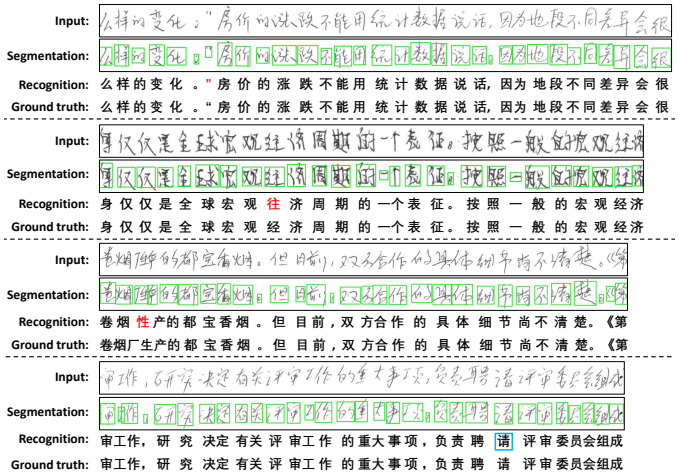


Fig. 9. Visualization results of ICDAR2013-Offline. The recognition results without language model are presented.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

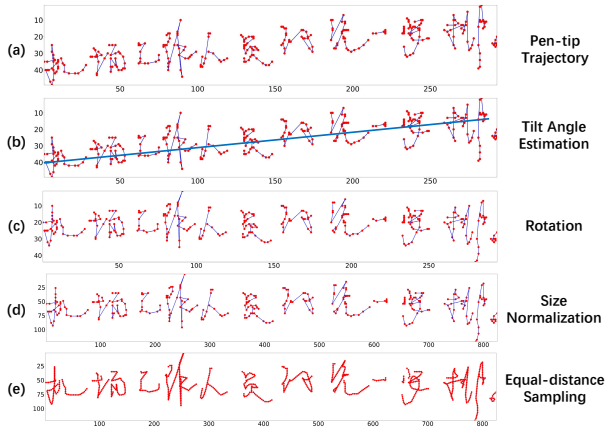


Fig. 10. Illustration of data preprocessing for the pen-tip trajectories from CASIA-OLHWDB2.0-2.2 and ICDAR2013-Online.

ICDAR2013-Online is depicted in Fig. 10. Given a pen-tip trajectory containing a sequence of points belonging to multiple strokes (Fig. 10(a)), we first perform linear fitting based on the (x, y) coordinates of the points and obtain the tilt angle of the text line, as shown in Fig. 10(b). Then, the text line is rotated to be horizontal, as depicted in Fig. 10(c). Next, the (x, y) coordinates of the points are rescaled to normalize the height of the text line to 128 while maintaining the aspect ratios, as shown in Fig. 10(d). Because the raw data adopts uniform-time sampling, the density of the points is related to the writing speed and sampling rate. Therefore, we resample the trajectory in an equal-distance manner as shown in Fig. 10(e). Specifically, the points are sampled at intervals of a Euclidean distance of 1.

For the synthetic pen-tip trajectories, only size normalization (Fig. 10(d)) and equal-distance sampling (Fig. 10(e)) are adopted.

3) Path Signature: For online HCTR, it is crucial to translate the pen-tip trajectory into offline feature maps while retaining most of the online information. The path signature [52], [53], [54], [55] has been verified to be effective in both online Chinese character recognition [56], [57], [58] and online HCTR [16], [18], [32]. Therefore, following previous methods [32], the truncated path signature feature maps up to 2nd integrated integral are calculated in a sliding-window fashion with a window size of 9. Fig. 11 visualizes the generated path signature feature maps.

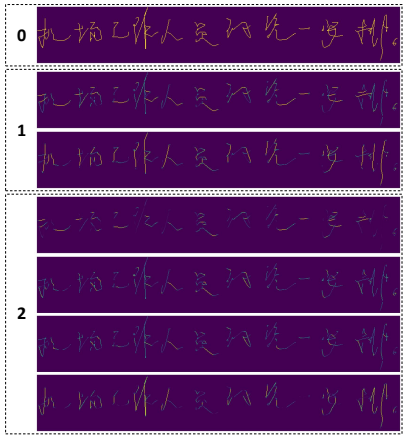


Fig. 11. Visualization of path signature feature maps (up to 2nd order).

TABLE II
COMPARISON WITH EXISTING METHODS ON ICDAR2013-ONLINE (LM: LANGUAGE MODEL)

Method	Without LM		With LM	
	AR	CR	AR	CR
Zhou et al. [2]	-	-	94.06	94.76
Zhou et al. [5]	-	-	94.22	94.76
Sun et al. [59]	89.12	90.18	93.40	94.43
2C-FCRN+impLM [32]	88.88	90.17	95.46	96.01
2C-FCRN+impLM&staLM [32]	88.88	90.17	96.06	96.58
VGG-DBLSTM [18]	87.49	87.98	97.03	97.29
CharNet-DBLSTM [18]	87.10	87.71	96.87	97.15
Liu et al. [19]	91.36	92.37	94.89	95.70
Peng et al. [64]	**95.05**	**95.46**	97.36	97.63
Ours (tri-gram LM)	94.46	94.67	96.64	97.28
Ours (Transformer-based LM)	94.46	94.67	**97.89**	**98.06**

4) Experimental Results: In Table II, we compare our methods with existing approaches on ICDAR2013-Online. When language models are not used, the performance of our method is comparable to state-of-the-art performance. Although the method proposed by [64] performs slightly better, the advantage of our method is that it can produce character segmentation results. When using language models, our approach equipped with the Transformer-based language model outperforms existing methods, including the method in [64] which also adopts a Transformer-based language model.

Fig. 12 illustrates the visualization results of ICDAR2013-Online. According to the predicted bounding boxes, we can divide the pen-tip trajectory into several segments, each of which corresponds to one character. Specifically, the points within a bounding box belong to the corresponding character. For the points that are not within any bounding boxes or are within multiple bounding boxes, they belong to the character corresponding to the closest bounding box. As shown in Fig. 12, both the segmentation and recognition results are very accurate.

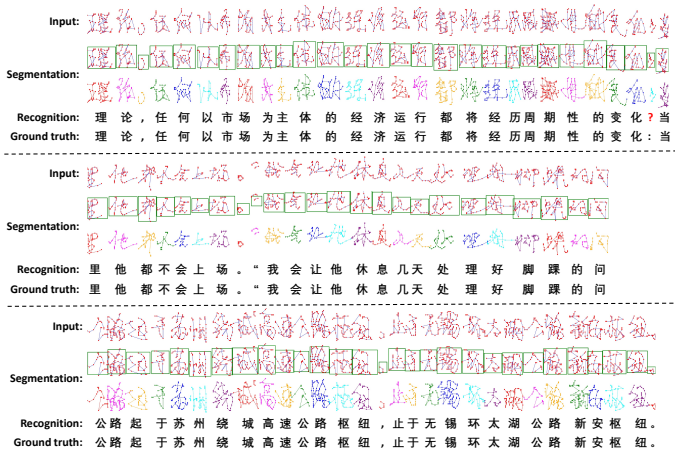


Fig. 12. Visualization results of ICDAR2013-Online. The recognition results without language model are presented.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

H. Experiments on SCUT-HCCDoc

1) Experimental Settings: First, the model is pretrained using the synthetic data (without semantic information) for SCUT-HCCDoc. Then, both the 93,411 real samples from the training set of SCUT-HCCDoc and the synthetic samples (without semantic information) for SCUT-HCCDoc are utilized to train the model. Because the vocabularies of SCUT-HCCDoc and the corpora described in Section IV-D are very different, the synthetic samples during the training stage are also synthesized without semantic information. Moreover, the ratios of the real and synthetic samples in a batch are 0.7 and 0.3, respectively. After the model training, the 23,218 samples from the testing set of SCUT-HCCDoc are adopted to evaluate the method. The number of character categories is set to 6,109.

2) Data Preprocessing: Both the real and synthetic text-line images are resized to a height of 128 pixels while maintaining their aspect ratios.

3) Experimental Results: In Table III, we compare our method with existing approaches on SCUT-HCCDoc. Specifically, the latest results of the CTC/attention-based approaches [14], [60], which were updated by the authors of [28] at their website[2], are presented in Table III. The performances of the other two methods [65], [30] are obtained from our reimplementation based on their official codes[3,4]. It can be observed that our method achieves state-of-the-art performance on SCUT-HCCDoc with an AR of 90.71% and a CR of 92.01%.

The visualization results of SCUT-HCCDoc are shown in Fig. 13. Although the synthetic samples for SCUT-HCCDoc are composed of simple characters generated from font files and white backgrounds as illustrated in Fig. 7(c), our method can still make full use of them and learn to segment and recognize characters of complex real samples. From the visualizations in Fig. 13, we can observe that our method can handle various writing styles including illegible handwriting and is unaffected by the noises such as illumination and the interference from the background. Especially for the sample at the top-right of Fig. 13, where the text is crossed out, the characters can still be successfully segmented and recognized, which verifies the robustness of our approach.

[2]https://github.com/HCIILAB/SCUT-HCCDoc_Dataset_Release
[3]https://github.com/intel/handwritten-chinese-ocr-samples
[4]https://github.com/Wang-Tianwei/Decoupled-attention-network

TABLE III
COMPARISON WITH EXISTING METHODS ON SCUT-HCCDOC

Method	AR	CR
CTC-based [14]	87.46	88.83
Attention-based [60]	83.30	84.81
DAN [30]	83.53	85.41
Liu et al. [65]	89.06	90.12
Ours	**90.71**	**92.01**

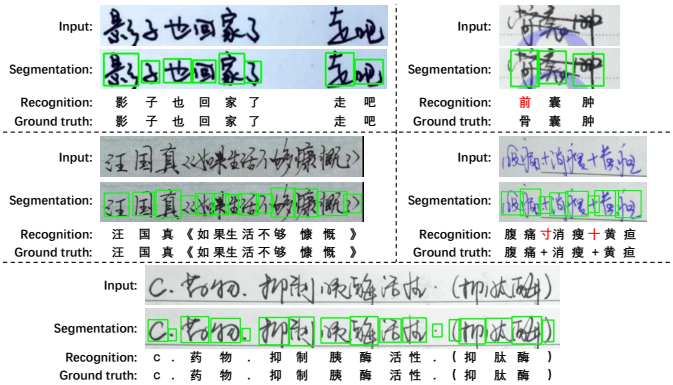


Fig. 13. Visualization results of SCUT-HCCDoc.

I. Ablation Studies

In Table IV, we conduct ablation analysis to demonstrate the effectiveness of each component of our method.

The baseline method represents the pretrained model that uses only synthetic data. It can be seen that training the network using only synthetic data leads to very poor performance. Especially for SCUT-HCCDoc where the real images are very different from the synthetic samples, the baseline can only achieve an extremely low accuracy with an AR of 1.14% and a CR of 1.25%.

When the proposed weakly supervised learning method is adopted, real samples with only transcript annotations can also be utilized to train the model. Taking advantage of the effective design of the weakly supervised learning method, the model can make full use of the general knowledge acquired from simple synthetic data and be adapted to handle the real sample by learning from its past successful predictions. Table IV shows that performance can be significantly improved. Even though the pretrained model for SCUT-HCCDoc has extremely poor performance, our weakly supervised learning method can also work very well.

ConR is aimed at integrating contextual information into the feature maps for character classification. The results in Table IV demonstrate that ConR can bring remarkable improvement.

J. Effectiveness of Weakly Supervised Learning

We further analyze the effectiveness of the proposed weakly supervised learning method. In our method, we distinguish the correct prediction by computing the edit distance between

TABLE IV
EFFECTIVENESS OF THE WEAKLY SUPERVISED LEARNING AND CONTEXTUAL REGULARIZATION

Method	ICDAR2013-Online		ICDAR2013-Offline		SCUT-HCCDoc	
	AR	CR	AR	CR	AR	CR
Baseline	56.18	56.90	59.68	60.38	1.14	1.25
+Weakly supervised learning	93.01	93.23	93.05	93.30	90.00	91.37
+Contextual regularization	**94.46**	**94.67**	**94.50**	**94.70**	**90.71**	**92.01**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3146771, IEEE Transactions on Multimedia

TABLE V
EFFECTIVENESS OF THE WEAKLY SUPERVISED LEARNING METHOD.

Method	ICDAR2013-Online		ICDAR2013-Offline		SCUT-HCCDoc	
	AR	CR	AR	CR	AR	CR
Text length-based [61], [62]	51.45	51.54	63.89	64.94	86.98	90.13
Ours	**93.01**	**93.23**	**93.05**	**93.30**	**90.00**	**91.37**

the recognition result and transcript annotation. Moreover, the pseudo bounding boxes are updated in a weighted-sum manner, considering all the correct predictions in previous iterations and their scores.

However, some existing approaches [61], [62] that involve weakly supervised learning follow a very simple pipeline. If the lengths of the recognition result and transcript annotation are equal, the predicted bounding boxes are directly viewed as the ground truth. Following their ideas, we replace our weakly supervised learning method with a new one named *Text Length*, where the pseudo bounding boxes B_{pse} are updated as

$$B_{pse} = \begin{cases} R_{seg}, & l_{pr} = l_{gt}, \\ B_{pse}, & otherwise, \end{cases} \tag{16}$$

where R_{seg}, l_{pr}, and l_{gt} are the segmentation result, the length of the recognition result, and the length of the transcript annotation, respectively (as specified in Section III-C).

As presented in Table V, our method outperforms the *Text length* by a large margin, especially for ICDAR2013-Online and ICDAR2013-Offline that contain long texts. All models in Table V are trained without contextual regularization. The criterion of *Text length* (i.e., the length of predictions and annotations are equal) cannot guarantee the quality of the predictions used for updating the pseudo bounding boxes, and is rarely satisfied for long texts and pretrained models with low accuracy. Moreover, the pseudo bounding boxes are updated by copying new predictions, which may make the training easily be interfered with by poor predictions. However, our method is based on the observation that correctly recognized characters are likely to have accurate bounding boxes. Thus only the characters that are matched as "equal" in computing edit distance are selected to ensure the quality of bounding boxes. Furthermore, the updating of pseudo bounding boxes is also carefully designed to suppress the impact of potential poor predictions.

By directly using the character bounding boxes provided by the annotations of CASIA-HWDB2.0-2.2, we can also train our model under full supervision. In Table VI, we compare the performance of our method on ICDAR2013-Offline under different supervisions, where the two models are trained with the same experimental settings as specified in Sections IV-C and IV-F. It can be seen that the weakly supervised model can even reach slightly higher AR and CR compared with the fully supervised counterpart, which verifies the effectiveness of our weakly supervised learning method. The superior performance under weak supervision may be due to the iterative pseudo bounding box updating mechanism. Specifically, the model is supervised by different pseudo bounding boxes every time perceiving the same text-line image, which may play a role of regularization.

TABLE VI
COMPARISON WITH FULLY SUPERVISED LEARNING ON OFFLINE SUBSET OF ICDAR2013 COMPETITION DATASET

Method	AR	CR
Fully supervised	94.43	94.64
Weakly supervised	**94.50**	**94.76**

TABLE VII
COMPARISON OF THE PERFORMANCE WITH AND WITHOUT THE BLSTM LAYERS. THE "RATIO" COLUMN PRESENTS THE RATIO OF THE TIME CONSUMED ON THE BLSTM LAYERS TO THE TIME CONSUMED ON THE ENTIRE NETWORK.

Dataset	Without BLSTM		With BLSTM		
	AR	CR	AR	CR	Ratio
ICDAR2013-Online	**94.46**	**94.67**	94.38	94.62	42%
ICDAR2013-Offline	**94.50**	**94.76**	94.46	94.72	42%
SCUT-HCCDoc	90.71	92.01	**90.85**	**92.11**	23%

K. Effectiveness of Contextual Regularization

In Table VII, experiments are conducted for further analysis of ConR.

First, we investigate the performance of the model if the BLSTM layers and the subsequent character classification results are adopted during inference, instead of using the character classification results from feature map f_{cls}. As shown in Table VII, the performances with and without the BLSTM layers are nearly the same. Specifically, compared with the counterpart with BLSTM layers, the AR and CR without BLSTM layers can be slightly higher on ICDAR2013-Online and ICDAR2013-Offline, and only drop a little on SCUT-HCCDoc. Based on the above results, we can conclude that ConR can really integrate the contextual information into the feature f_{cls} before BLSTM layers.

As for the inference speed, the two BLSTM layers can consume more than 40% of the inference time of the entire network if they are used for context modeling, as demonstrated by the "Ratio" column of Table VII. This is attributed to the non-parallel running of the BLSTM layer and the long texts that are common in Chinese documents. With the help of ConR, the recurrent layers are not required during inference, thus significantly improving the speed.

L. Comparison with CTC and Attention

Table VIII compares our method with existing widely used CTC/attention-based approaches.

For ICDAR2013-Online and ICDAR2013-Offline, two representative CTC-based methods [19], [21] that follow mainstream CNN+RNN+CTC or RNN+CTC architectures are

TABLE VIII
COMPARISON WITH CTC-BASED AND ATTENTION-BASED METHODS

Dataset	Method	AR	CR	Speed
ICDAR2013-Online	CTC-based [19]	91.36	92.37	62fps
	Attention-based [36]	85.35	85.84	16fps
	Ours	**94.46**	**94.67**	**70fps**
ICDAR2013-Offline	CTC-based [21]	91.55	92.13	64fps
	Attention-based [36]	84.79	85.90	16fps
	Ours	**94.50**	**94.70**	**70fps**
SCUT-HCCDoc	CTC-based [14]	87.64	88.83	75fps
	Attention-based [60]	83.30	84.81	52fps
	Ours	**90.71**	**92.01**	**97fps**

adopted for comparison. Owing to the lack of attention-based methods in previous literature, we reimplement the sequence recognition network of [36] on ICDAR2013-Online and ICDAR2013-Offline with the same training data and preprocessing as ours. Because the network [36] is designed for offline images, the path signature feature maps (Section IV-G3) are also adopted for online texts. For SCUT-HCCDoc, we compare our method with CTC/attention-based approaches [14], [60] adopted in [28] (same as Table III). Furthermore, we also test the inference speed of these methods and ours using an NVIDIA GTX 1080ti GPU with 11GB of memory.

The results in Table VIII show that our method outperforms existing CTC/attention-based approaches in terms of AR and CR. Moreover, owing to the parallel computing characteristic of the fully convolutional architecture, our method exhibits a higher inference speed compared with the CTC/attention-based methods that use recurrent layers for contextual modeling. In addition, our method can produce character segmentation results, whereas CTC/attention-based approaches cannot.

M. Reading Chinese Text in the Wild

In this section, we further explore the potential of our method to be extended to the recognition of Chinese text in the wild. The experiments are conducted using ReCTS-25k [34], which contains 25,000 signboard images (20,000 for training and 5,000 for testing). There are 108,963 and 10,789 text lines cropped from the training and testing sets, respectively. We further divide the text lines from the training set into 90,763 samples for training and 18,200 samples for validating. The synthetic data is synthesized in the same way as in Section IV-B, using white background and characters from font files. The number of character categories is 4,134. Other details follow the experiments on SCUT-HCCDoc and Section IV-C.

As shown in Table IX, our method achieves a normalized edit distance (NED) [34] of 86.66% on the validating set and 90.70% on the testing set. Some visualization results are shown in Fig. 14. When tested using an NVIDIA GTX 1080ti GPU, our method performs better with a higher inference speed. However, our approach is inferior to the winning method (SNAHL_v1 with 95.55% NED) of the competition [34], which uses an ensemble of three types of models and a large amount of external data. To conclude, our method has the potential of being extended to other applications besides online and offline HCTR.

TABLE IX
PERFORMANCE ON RECTS-25K DATASET

Method	Validating NED	Testing NED	Speed
CTC-based [14]	80.21	83.40	86fps
Attention-based [36]	82.02	86.46	69fps
SANHL_v1*[34]	-	**95.55**	-
Ours	**86.66**	90.70	**108fps**

* The winner of the ICDAR 2019 ReCTS competition [34].

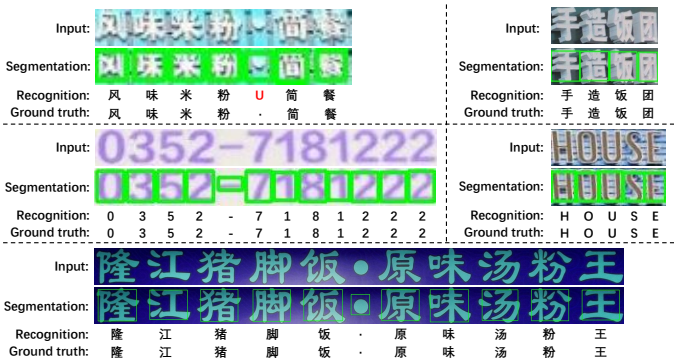


Fig. 14. Visualization results of ReCTS-25k.

N. Error Analysis

The weakness of our method lies in the segmentation of punctuations. Fig. 15 presents the failure cases of punctuation segmentation (indicated by red arrows). Compared with Chinese characters, punctuations are much smaller and have more flexible locations that could be the top or bottom of the text line. However, punctuations are overwhelmed by Chinese characters in the training data, which may be the major cause of the poor segmentation of punctuations. Nevertheless, owing to the decoupled design of character segmentation and classification, the recognition results can still be correctly predicted in the first two cases of Fig. 15. Moreover, the punctuation could be very close to its previous character especially for handwritten texts, which makes it difficult to distinguish. For example, the comma in the third case of Fig. 15 is omitted by our method.

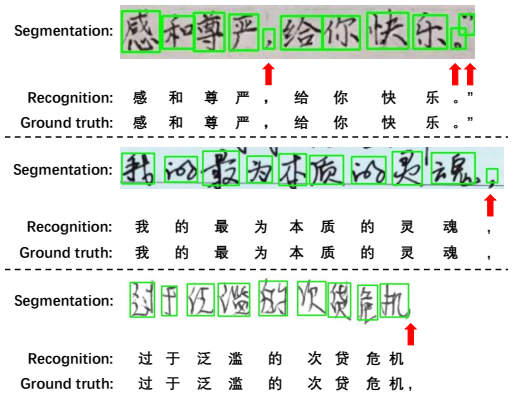


Fig. 15. Failure cases of the segmentation of punctuations.

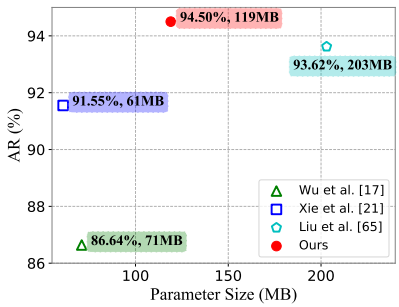


Fig. 16. AR versus parameter size on ICDAR2013-Offline.

O. Discussion

1) Parameter Size: Our method is basically for offline text recognition because online texts are also expressed as offline representations. Thus, we illustrate the AR versus parameter size of existing methods on ICDAR2013-Offline in Fig. 16. Note that only methods whose parameter sizes were reported in previous literature are presented. It can be observed that our model has a good trade-off between AR and parameter size. Especially compared with the method in [65], our method achieves better performance with approximately half of the parameters.

2) Training Time: As described in Section IV-C, there are two stages of our model training, i.e., the pretraining and training stages. Using an NVIDIA RTX 2080ti GPU, it takes approximately 125 hours in total to train a model for ICDAR2013-Offline. Specifically, the pretraining and training stages take 14 hours and 111 hours, respectively. The time required to train a model for ICDAR2013-Online is nearly the same because we translate the online pen-tip trajectories into image-like representations. In previous literature, there are few studies which reported their training time. For example, the method in [65] for ICDAR2013-Offline requires approximately 80 hours, and the method in [19] for ICDAR2013-Online requires 102 hours. Compared with these methods, the training time of our method is slightly longer but still acceptable. Because most procedures of the weakly supervised learning run on CPU in our implementation, the training speed could be accelerated by migrating them to GPU in the future.

3) Failure Case of Weakly Supervised Learning: If the ground truth consists of multiple same characters, such as "AAA", but only "AA" is recognized, we can not determine which two "A"s in the ground truth are correctly recognized by calculating edit distance. This issue could lead to inaccurate pseudo bounding boxes because they may be updated using unsuitable predictions. However, such a situation rarely occurs in real data that contains natural texts. Thus, the model training is almost unaffected. Nevertheless, this problem can be explored by incorporating spatial constraints or feature similarities in the future.

V. Conclusion

In this paper, we propose a novel segmentation-based method for online and offline HCTR. In contrast to previous oversegmentation-based approaches, we formulate a brand-new segmentation-based text recognition framework that end-to-end segments and recognizes characters through fully convolutional networks with high efficiency and accuracy. To address the high cost of character segmentation annotations, a new weakly supervised learning method is proposed to enable the network to be trained using only transcript annotations. Owing to the absence of context modeling in the fully convolutional architecture, we design a contextual regularization method to integrate contextual information into extracted features without affecting the inference speed. Extensive experiments on CASIA-HWDB, CASIA-OLHWDB, ICDAR2013, and SCUT-HCCDoc, demonstrate the superiority of our method over existing approaches. To the best of our knowledge, our method may be the first to achieve state-of-the-art performance on both online and offline HCTR. An additional trial on ReCTS-25k demonstrates the potential of our method out of the field of HCTR. We hope this work will spark further research beyond the realms of prevalent CTC/attention-based methods.

Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003, 61771199), and GD-NSF (no. 2017A030312006).

References

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

[2] X. Zhou, D. Wang, F. Tian, C. Liu, and M. Nakagawa, "Handwritten Chinese/Japanese text recognition using semi-Markov conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2413–2426, 2013.

[3] 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.

[4] F. Yin, Q. Wang, X. Zhang, and C. Liu, "ICDAR 2013 Chinese handwriting recognition competition," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2013, pp. 1464–1470.

[5] X.-D. Zhou, Y.-M. Zhang, F. Tian, H.-A. Wang, and C.-L. Liu, "Minimum-risk training for semi-Markov conditional random fields with application to handwritten Chinese/Japanese text recognition," *Pattern Recognit.*, vol. 47, no. 5, pp. 1904–1916, 2014.

[6] 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. Int. Conf. Front. Handwrit. Recognit.*, 2016, pp. 84–89.

[7] 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, 2017.

[8] Z.-X. Wang, Q.-F. Wang, F. Yin, and C.-L. Liu, "Weakly supervised learning for over-segmentation based handwritten Chinese text recognition," in *Proc. Int. Conf. Front. Handwrit. Recognit.*, 2020, pp. 157–162.

[9] 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, 2009.

[10] J. Du, Zi-Rui Wang, J. Zhai, and J. Hu, "Deep neural network based hidden Markov model for offline handwritten Chinese text recognition," in *Proc. Int. Conf. Pattern Recognit.*, 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. Doc. Anal. Recognit.*, vol. 21, no. 4, pp. 241–251, 2018.

[12] 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, p. 107102, 2020.

[13] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, 2009.

[14] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017.

[15] R. Messina and J. Louradour, "Segmentation-free handwritten Chinese text recognition with LSTM-RNN," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2015, pp. 171–175.

[16] Z. Xie, Z. Sun, L. Jin, Z. Feng, and S. Zhang, "Fully convolutional recurrent network for handwritten Chinese text recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 4011–4016.

[17] Y. Wu, F. Yin, Z. Chen, and C. Liu, "Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2017, pp. 79–84.

[18] K. Chen, L. Tian, H. Ding, M. Cai, L. Sun, S. Liang, and Q. Huo, "A compact CNN-DBLSTM based character model for online handwritten Chinese text recognition," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2017, pp. 1068–1073.

[19] M. Liu, Z. Xie, Y. Huang, L. Jin, and W. Zhou, "Distilling GRU with data augmentation for unconstrained handwritten text recognition," in *Proc. Int. Conf. Front. Handwrit. Recognit.*, 2018, pp. 56–61.

[20] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, and L. Xie, "Aggregation cross-entropy for sequence recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6538–6547.

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

[22] Z.-Y. Zhu, F. Yin, and D.-H. Wang, "Attention combination of sequence models for handwritten Chinese text recognition," in *Proc. Int. Conf. Front. Handwrit. Recognit.*, 2020, pp. 288–294.

[23] 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, 2018.

[24] J. Zhang, J. Du, Y. Yang, Y.-Z. Song, and L. Dai, "SRD: A tree structure based decoder for online handwritten mathematical expression recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 2471–2480, 2021.

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Machin. Learn.*, 2006, pp. 369–376.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[27] U.-V. Marti and H. Bunke, "The IAM-database: an english sentence database for offline handwriting recognition," *Int. J. Doc. Anal. Recognit.*, vol. 5, no. 1, pp. 39–46, 2002.

[28] H. Zhang, L. Liang, and L. Jin, "SCUT-HCCDoc: A new benchmark dataset of handwritten Chinese text in unconstrained camera-captured documents," *Pattern Recognit.*, vol. 108, p. 107559, 2020.

[29] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 5076–5084.

[30] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, "Decoupled attention network for text recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12 216–12 224.

[31] 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. Doc. Anal. Recognit.*, 2019, pp. 25–30.

[32] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten Chinese text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1903–1917, 2018.

[33] C. Liu, F. Yin, D. Wang, and Q. Wang, "CASIA online and offline Chinese handwriting databases," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2011, pp. 37–41.

[34] R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, X. Bai, B. Shi, D. Karatzas, S. Lu, and C. V. Jawahar, "ICDAR 2019 robust reading challenge on reading Chinese text on signboard," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2019, pp. 1577–1581.

[35] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: An attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, 2019.

[36] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4168–4176.

[37] H. Wu, X. Ma, and Y. Li, "Convolutional networks with channel and STIPs attention model for action recognition in videos," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2293–2306, 2020.

[38] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2990–3001, 2020.

[39] N. Zhao, H. Zhang, R. Hong, M. Wang, and T.-S. Chua, "VideoWhisper: Toward discriminative unsupervised video feature learning with attention-based recurrent neural networks," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2080–2092, 2017.

[40] Y. Xiu, Q. Wang, H. Zhan, M. Lan, and Y. Lu, "A handwritten Chinese text recognizer applying multi-level multimodal fusion network," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2019, pp. 1464–1469.

[41] X. Zhang, F. Yin, Y. Zhang, C. 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, 2018.

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

[43] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 850–855.

[44] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," in *Adv. Neural Inform. Process. Syst. Deep Learn. Workshop*, 2014.

[45] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2315–2324.

[46] F. Zhan, S. Lu, and C. Xue, "Verisimilar image synthesis for accurate detection and recognition of texts in scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 249–266.

[47] S. Fogel, H. Averbuch-Elor, S. Cohen, S. Mazor, and R. Litman, "ScrabbleGAN: Semi-supervised varying length handwritten text generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 4324–4333.

[48] "The people's daily corpus," [Online]. Available: http://icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp, the People's Daily News and Information Center, the Peking University Institute of Computational Linguistics and Fujitsu Research and Development Center Limited. Accessed on: Mar. 25, 2016.

[49] G. Jin, "The PH corpus," [Online]. Available: ftp://ftp.cogsci.ed.ac.uk/pub/chinese, accessed on: Mar. 25, 2016.

[50] Chinese linguistic data consortium, [Online]. Available: http://www.chineseldc.org, 2009, the Contemporary Corpus developed by State Language Commission P.R.China, Institute of Applied Linguistics, Accessed on: Oct. 22, 2016.

[51] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Machin. Learn.*, 2014, pp. 1764–1772.

[52] K.-T. Chen, "Integration of paths–a faithful representation of paths by noncommutative formal power series," *Trans. Amer. Math. Soc.*, vol. 89, no. 2, pp. 395–407, 1958.

[53] T. Lyons, Z. Qian, Z. Qian *et al.*, *System control and rough paths*. Oxford, UK: Clarendon, 2002.

[54] T. Lyons, "Rough paths, signatures and the modelling of functions on streams," *arXiv preprint arXiv:1405.4537*, 2014.

[55] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Ann. Math.*, pp. 109–167, 2010.

[56] B. Graham, "Sparse arrays of signatures for online character recognition," *arXiv preprint arXiv:1308.0371*, 2013.

[57] W. Yang, L. Jin, Z. Xie, and Z. Feng, "Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2015, pp. 551–555.

[58] 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.

[59] L. Sun, T. Su, C. Liu, and R. Wang, "Deep LSTM networks for online Chinese handwriting recognition," in *Proc. Int. Conf. Front. Handwrit. Recognit.*, 2016, pp. 271–276.

[60] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 577–585.
[61] L. Xing, Z. Tian, W. Huang, and M. R. Scott, "Convolutional character networks," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9126–9136.
[62] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 9365–9374.
[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.
[64] D. Peng, C. Xie, H. Li, L. Jin, Z. Xie, K. Ding, Y. Huang, and Y. Wu, "Towards fast, accurate and compact online handwritten Chinese text recognition," in *Proc. Int. Conf. Doc. Anal. Recognit.*, 2021, pp. 157–171.
[65] B. Liu, W. Sun, W. Kang, and X. Xu, "Searching from the prediction of visual and language model for handwritten Chinese text recognition," in *Proc. Int. Conf. Doc. Anal. Recognit*, 2021, pp. 274–288.
[66] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "Fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. Conf. N. Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019.
[67] C. Luo, L. Jin, and Z. Sun, "MORAN: A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, 2019.

Dezhi Peng received the B.S. degree in information engineering from South China University of Technology in 2019. He is currently pursuing the Ph.D. degree in information and communication engineering at South China University of Technology. His research interests include optical character recognition, document analysis and recognition, and handwriting text recognition.

Lianwen Jin received the B.S. degree from the University of Science and Technology of China, Anhui, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1991 and 1996, respectively. He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology. He is the author of more than 100 scientific papers. Dr. Jin was a recipient of the award of New Century Excellent Talent Program of MOE in 2006 and the Guangdong Pearl River Distinguished Professor Award in 2011. His research interests include computer vision, optical character recognition, handwriting analysis and recognition, machine learning, deep learning, and intelligent systems.

Weihong Ma received the B.S. degree from the school of Electronic and Information Engineering at the South China University of Technology, Guangzhou, China in 2019. He is currently pursuing the master degree in information and communication engineering at the South China University of Technology, Guangzhou, China. His current research interests include deep learning, scene text detection, and document analysis.

Canyu Xie received the B.S. degree in electronic science and technology at the South China University of Technology, Guangzhou, China in 2019. He is currently pursuing the master degree in electronic and communication engineering at South China University of Technology, Guangzhou, China. His current research interests include computer vision, model compression, and acceleration.

Hesuo Zhang received the B.S. degree from the school of Electronic and Information Engineering at South China University of Technology, Guangzhou, China in 2019. He is currently pursuing the master degree in information and communication engineering at South China University of Technology, Guangzhou, China. His current research interests include machine learning, deep learning, and handwritten text segmentation and recognition.

Shenggao Zhu received the Ph.D. degree in Computer Science from National University of Singapore (NUS), 2017. He got his bachelor in Electronic Engineering and Information Science from University of Science and Technology of China (USTC), 2011. He joined Huawei Cloud in 2017 and now is a Technical Expert. His research interests include computer vision and AI applications.

Jing Li received the B.S. degree from the University of Science and Technology of China, and the Ph.D. degree from the National University of Singapore, in 2013 and 2019, respectively. She is now an AI engineer in Huawei Cloud Computing Technologies. Her research interests include face recognition, optical character recognition and image generation.