IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 18, 2018, accepted November 30, 2018, date of publication December 6, 2018,
date of current version January 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885398

SCUT-EPT: New Dataset and Benchmark for Offline Chinese Text Recognition in Examination Paper

YUANZHI ZHU[1], ZECHENG XIE[1], LIANWEN JIN[1], (Member, IEEE), XIAOXUE CHEN[1], YAOXIONG HUANG[1], AND MING ZHANG[2]

[1]College of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
[2]AbcPen Inc., Hangzhou, China

Corresponding author: Lianwen Jin (lianwen.jin@gmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001405 and Grant GD-NSF 2017A030312006, in part by NSFC under Grant 61673182 and Grant 61771199, and in part by GDSTP under Grant 2017A010101027 and Grant GZSTP 201607010227.

ABSTRACT Most existing studies and public datasets for handwritten Chinese text recognition are based on the regular documents with clean and blank background, lacking research reports for handwritten text recognition on challenging areas such as educational documents and financial bills. In this paper, we focus on examination paper text recognition and construct a challenging dataset named examination paper text (SCUT-EPT) dataset, which contains 50 000 text line images (40 000 for training and 10 000 for testing) selected from the examination papers of 2 986 volunteers. The proposed SCUT-EPT dataset presents numerous novel challenges, including character erasure, text line supplement, character/phrase switching, noised background, nonuniform word size, and unbalanced text length. In our experiments, the current advanced text recognition methods, such as convolutional recurrent neural network (CRNN) exhibits poor performance on the proposed SCUT-EPT dataset, proving the challenge and significance of the dataset. Nevertheless, through visualizing and error analysis, we observe that humans can avoid vast majority of the error predictions, which reveal the limitations and drawbacks of the current methods for handwritten Chinese text recognition (HCTR). Finally, three popular sequence transcription methods, connectionist temporal classification (CTC), attention mechanism, and cascaded attention-CTC are investigated for HCTR problem. It is interesting to observe that although the attention mechanism has been proved to be very effective in English scene text recognition, its performance is far inferior to the CTC method in the case of HCTR with large-scale character set.

INDEX TERMS Offline handwritten Chinese text recognition (HCTR), educational documents, sequence transcription.

I. INTRODUCTION

Handwriting recognition of different languages are challenging issues and receive extensive attention from researchers. In recent years, numerous handwritten datasets have been published in the field to promote the advancement of the community. In general, handwritten datasets can be divided into two categories, i.e., online and offline datasets. For example, there are offline handwritten datasets such as French paragraph dataset Rimes [2], English text dataset IAM [3], Arabic datasets of IFN/ENIT [4] and KHATT [5], Chinese dataset CASIA-HWDB [6] and HIT-MW [7]. For online handwritten datasets, there are Japanese text datasets Kondate [8] and character dataset TUAT Nakayosi_t and Kuchibue_d [9], English text dataset IAM-OnDB [10], Chinese datasets SCUT-COUCH2009 [11], CASIA-OLHWDB [6], and ICDAR2013 competition set [12]. Specially, Chinese handwriting recognition has the challenges of handwritten styles diversity, mis-segmentation, and large-scale character set, and attracts a large number of researchers [13]–[15]. Generally, Chinese handwriting recognition can be divided into four categories [6]: online/offline handwritten character/text recognition. However, with the recent rapid development of deep learning technology, researchers have pushed the recognition performance to a fairly high level, e.g., 96.28 of correct rate for offline Chinese text recognition on the test set of CASIA-HWDB [16]. Such a high recognition result suggests that the main recognition problems associated with existing popular offline Chinese text dataset, e.g. CASIA-HWDB 2.0-2.2 [6], have been basically solved. In other words, the community desires more complicate and challenging datasets for performance evaluation of the latest technologies on handwriting recognition.

Writing style diversity and large-scale character set [17], [18] are fundamental issues in traditional handwritten Chinese text recognition [6]. Conventionally, integrated segmentation-recognition method [13], [19] constructs the segmentation-recognition lattice based on sequential character segments of text line images, followed by optimal path searching by integrating the recognition scores, geometry information, and semantic context, but may suffer from the problem of mis-segmentation [20], [21]. Recently, the combination of convolutional neural network (CNN) and long short-term memory (LSTM) [22] exhibits excellent performance in the fields such as scene text recognition [1], [23], handwritten text recognition [24], [25] and action and gesture recognition [26], [27]. Fully convolutional recurrent network [24] and its improved architecture multi-spatial-context fully convolution recurrent network [14] are one of the existing state-of-the-art text recognition frameworks for online handwritten Chinese text recognition problems. Specifically, the above-described deep learning based networks primarily apply Connectionist Temporal Classification (CTC) decoder [28] for end-to-end sequential training, completely avoiding explicit alignment between input images and their corresponding label sequences. Another transcription method, attention mechanism, is popular in machine translation [29] for unfixed-order transcription between different languages, and is successfully applied in scene text recognition [30], [31] problem with state-of-the-art performance. Recently, a new method that combines attention mechanism and CTC achieved state-of-the-art result on the field of lipreading [32] and speech recognition [33], [34]. Specifically, Kim *et al.* [34] use CTC objective function as an auxiliary task to train the attention model encoder within the multi-task learning (MTL) framework. In contrast, Xu *et al.* [32] and Das *et al.* [33] directly incorporating attention within the CTC framework, namely cascaded attention-CTC decoder in this paper. However, to the best of our knowledge, both the attention mechanism and cascaded attention-CTC decoder have not yet made breakthrough progress in handwritten Chinese text recognition problem.

In this paper, we present an offline text recognition dataset, named Examination Paper Text (SCUT-EPT)[1] dataset, for examination paper text recognition in the education field. The proposed SCUT-EPT Dataset contains 50,000 text line images, including 40,000 for training and 10,000 for testing, selected from examination papers of 2,986 volunteers. In addition to the common problems in HCTR, Dataset SCUT-EPT also encounters novel challenges in examination paper, including character erasure, text line supplement, character/phrase switching, noised background, nonuniform word size and unbalanced text length, as shown in Fig. 4. *Character erasure*, also known as crossed-outs [35]–[37], often accompanies with crossed lines to strike out characters; *Text line supplement* occurs with additional text line supplement appearing below or above the normal text line; *Character/phrase switching* is the phenomenon where writers add special symbols to switch relevant written characters or phrases for better understanding; *Noised background* refers to underlines below characters, dense grids between characters, etc. in contrast to most of the handwritten datasets [3], [6], [7], [11] whose backgrounds are very clean; *Nonuniform word size* refers to the nonuniform word size of characters, especially when comparing Chinese character with digit, letter and symbol; *Unbalanced text length* usually comes from different types of questions that result in different length of answers in the exam papers.

In the experiments, we evaluate the state-of-the-art recognition method CRNN [1] on the proposed dataset and observe poor performance. However, visualization shows that majority of the error recognized images can be correctly recognized by human eyes, but easily confused by current mainstream recognition methods, which exposes the limitations of existing text recognition technology. Considering the difficulty of the dataset, we make a comprehensive investigation on CTC, attention mechanism and cascaded attention-CTC for HCTR problem. It is worth noting that although attention mechanism has shown promising performance in scene text recognition of western language [23], [30], [38], it fail to provide acceptable result for HCTR problem. In the experiment, we found that CTC-based seq-to-seq method exhibits superior performance over attention and cascaded attention-CTC on dataset SCUT-ETP. Specifically, the proposed solution in this paper for dataset SCUT-EPT consists of three components, including fully convolutional network for feature extraction, multi-layered residual LSTM [14] for context learning, and CTC for transcription.

Overall, the novel contributions this paper offers can be summarized as follow:

1) A new large-scale offline handwritten Chinese text dataset named SCUT-EPT with numerous novel challenges is presented to the community.
2) We present baseline experiments with the advanced recognition architecture, CRNN, on the proposed SCUT-EPT dataset, and provide detailed analysis to its poor recognition results.
3) This is the first work to compare the role of three popular sequence learning methods, i.e., CTC, attention mechanism and cascaded attention-CTC decoder for HCTR problem.

[1] Dataset SCUT-EPT is available at https://github.com/HCIILAB/SCUT-EPT_Dataset_Release.

TABLE 1. Detailed comparison of typical handwritten datasets in different language. The first part ([2]–[5], [8]–[10]) describe popular handwritten datasets of different languages, except Chinese. The second part ([6], [11], [12]) presents standard handwritten Chinese character datasets. The third part ([6], [7], [12] and SCUT-EPT) shows typical handwritten Chinese text datasets and the proposed SCUT-EPT dataset. Compared with other Chinese text datasets, the proposed SCUT-EPT not only has rich text lines and character samples but also possesses the most writers and classes.

Dataset name		Language	Online/Offline	Type	Writers	Text lines	Samples	Classes	Year
Rimes [2]		French	Offline	text	1,300	12,111	-	-	2008
IAM [3]		English	Offline	text	400	9,285	82,227	81	2002
IAM-OnDB [10]		English	Online	text	221	13,049	86,272	81	2005
IFN/ENIT [4]		Arabic	Offline	character	946	-	26,459	-	2002
KHATT [5]		Arabic	Offline	text	1000	-	-	49	2012
TUAT Kuchibue_d [9]		Japanese	Online	character	120	-	1,435,440	3356	2004
TUAT Nakayosi_t [9]		Japanese	Online	character	163	-	1,695,689	4438	2004
Kondate [8]		Japanese	Online	text	100	12,232	130,956	1106	2014
SCUT-COUCH2009 [11]		Chinese	Online	character	190	-	3,670,805	44,208	2009
CASIA-HWDB1.0-1.2 [6]			Offline		1020	-	3,895,135	7,356	2011
CASIA-OLHWDB1.0-1.2 [6]			Online		1020	-	3,912,017	7,356	2011
ICDAR2013 competition set [12]			Offline		60	-	224,419	3,755	2013
ICDAR2013 competition set [12]			Online		60	-	224,590	3,755	2013
HIT-MW [7]		Chinese	Offline	text	780	8,664	186,444	3,041	2006
CASIA-HWDB2.0-2.2 [6]			Offline		1,019	52,230	1,349,414	2,703	2011
CASIA-OLHWDB2.0-2.2 [6]			Online		1,019	52,221	1,348,904	2,655	2011
ICDAR2013 competition set [12]			Offline		60	3,432	91,563	1,385	2013
ICDAR2013 competition set [12]			Online		60	3,432	91,576	1,375	2013
SCUT-EPT	**Train**		**Offline**		**2986**	**40,000**	**1,018,432**	**4,058**	**2018**
	Test		**Offline**			**10,000**	**248,729**	**3,236**	
	Total		**Offline**			**50,000**	**1,267,161**	**4,250**	

The rest of this paper is organized as follows: Section 2 reviews existing handwritten datasets. Section 3 formally introduces the proposed dataset, its challenges, and its annotation methods in detail. Section 4 describes three transcription methods, including CTC decoder, attention decoder, and cascaded attention-CTC decoder. Section 5 presents the experimental results and analysis. Section 6 shows the conclusion and future work.

II. EXISTING HANDWRITTEN DATASETS

Since the twenty-first century, the document analysis and recognition community has published massive amount of new handwritten datasets in different languages for handwritten recognition studies, as shown in the first part of Table 1. Rimes database [2] is an offline French paragraph (text) dataset with a total of 1,600 paragraphs (12,111 text lines) contributed by 1300 writers. For English text recognition, IAM database [3] is an offline dataset consisting of 9,285 text lines (82,227 words) produced by approximately 400 writers, while IAM-OnDB [10] is an online dataset with a total of 13,049 text lines (86,272 words) from 221 writers. For offline Arabic words recognition, Pechwitz and Margner provided IFN/ENIT database [4] with a total of 26,459 handwritten words of 946 Tunisian town/villages names written by different writers. Another offline Arabic text database KHATT [5] consists of 1,000 handwritten forms written by 1,000 writers from different countries, which can be used for paragraph and line level recognition tasks. For online Japanese character recognition, Nakagawa and Matsumoto [9] proposed two important datasets, TUAT Nakayosi_t and Kuchibue_d, containing over three million patterns: one with 120 people contributing 11,962 patterns each and another with 163 participants contributing 10,403 patterns each. These two datasets store totally three million of characters mostly in text, with less frequently used characters collected character by character. As for online Japanese text recognition, Kondate database [8] with a total of 12,232 text lines collected from 100 people was contributed to the research community.

In the field of handwritten Chinese character recognition, SCUT-COUCH2009 database [11] is a comprehensive online unconstrained character database with totally 3.6 million character samples contributed by more than 190 persons. It consists of 11 datasets of isolated characters (Chinese simplified and traditional, English letters, digits, symbols), Chinese Pinyin and words. CASIA-HWDB1.0-1.2/CASIA-OLHWDB1.0-1.2 [6] are the currently existing most popular and comprehensive handwritten datasets for Chinese online/offline isolated character recognition evaluation, containing about 3.9 million samples of 7,356 classes (7,185 Chinese characters and 171 symbols). ICDAR2013 competition set (isolated characters) [12], collected for the evaluation of 2013 Chinese handwriting recognition competition, has both online and offline data for isolated character recognition. More details are summarized in the second part of Table 1.

For handwritten Chinese text datasets, the third part of Table 1 provides detailed comparisons between existing datasets with the proposed SCUT-EPT. In the early 2006, Su *et al.* [7] put forward the first handwritten Chinese text dataset, HIT-MW, including 8,664 text lines, totally 186,444 characters of 3,041 classes. HIT-MW is collected by mail or middleman instead of face to face, resulting in some real handwriting phenomena for academic research, such as miswriting and erasing. CASIA-HWDB2.0-2.2 and CASIA-OLHWDB2.0-2.2 [6] are large-scale datasets containing 1.35 million character samples (52.2 thousand text lines), with only 1019 writers and fewer than 3,000 categories. Either online or offline, the scale of ICDAR2013 competition set (continuous texts) [12] is smaller, with only 60 writers, 91.5 thousand character samples and less than 1400 classes. With the rapid development of deep learning technology, these datasets are no longer challenging or complicate enough to properly evaluate the latest technologies for HCTR problem. For example, state-of-the-art model achieves 96.28% (93.24%) correct rate on the testing set of CASIA-HWDB2.0-2.2 [16] with (without) language model, and Wu *et al.* [13] obtains the current highest 96.32% correct rate on the ICDAR2013 competition set.

To the best of our knowledge, existing researches and public datasets are mainly developed for handwritten text recognition on the regular document with clean background, lacking in research reports on handwritten text recognition of specific and challenging areas such as educational documents, financial bills. In this paper, the proposed SCUT-EPT dataset contains numerous novel challenges, such as character erasure, text line supplement, character/phrase switching, noised background, nonuniform word size and unbalanced text length. The above-mentioned challenges exist not only in educational documents, but also in paper letters, notebooks, handwritten receipts, financial bills, etc. Compared with traditional offline handwritten Chinese text datasets, the proposed dataset is more representative academic research of handwriting Chinese recognition in our life, which can better evaluate the most advanced recognition technologies and catalyze the emergence of new technologies.

In summary, compared with existing datasets, the advantages of the proposed SCUT-EPT are following:

1) SCUT-EPT is a large-scale dataset containing 1.26 million character samples (50,000 text line images), which is comparable to that of CASIA-HWDB2.0-2.2 [6], but far exceeds ICDAR2013 competition set [12] and HIT-MW [7].
2) Compared with other datasets, SCUT-EPT possess the most classes of 4,250 and writers of 2,986, significantly guaranteeing its diversity and richness.
3) Compared with other datasets, SCUT-EPT dataset is more relevant to the daily life with various challenges. Therefore, SCUT-EPT dataset is of vital importance to academic research and the evaluation of the latest recognition technologies.

III. EXAMINATION PAPER TEXT DATASET

In order to construct the SCUT-EPT dataset for examination papers, 2986 high school students are incorporated in this project to finish an examination paper. For privacy reasons, we only choose part of the text line images from the examination paper of each student and construct the SCUT-EPT dataset. The developed dataset contains 50,000 text images, including 40,000 text line images as training set and 10,000 text line images as testing set.

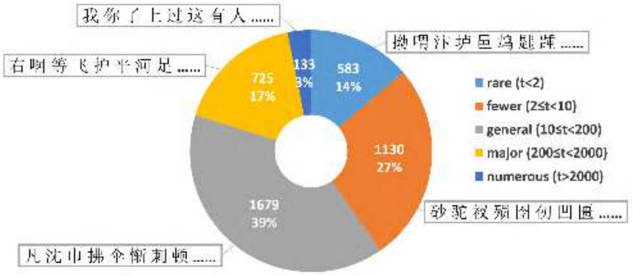

FIGURE 1. **The class distribution and typical samples of each grade (*t* represents number of character occurrence in SCUT-EPT).**

A. DATASET DESCRIPTION

As shown in Table 1, there are totally 4,250 classes in our dataset SCUT-EPT, including 4,033 commonly used Chinese characters, 104 symbols, and 113 outlier Chinese characters, where outlier Chinese character means that the Chinese character is outside the character set of the popular CASIA-HWDB1.0-1.2 [6]. It should be noted that there is no intersection between the training set and the testing set, i.e., students who contribute to the training set will not play a part in the testing set. The total character samples in the SCUT-EPT dataset is 1,267,161, with approximately 25 characters each text line.

In Fig. 1, we provide the class distribution as well as typical samples of each grade. It is clear that the class distribution is extremely unbalanced, classes with 10 or fewer samples occupy a proportion of 41% while 3% of classes has more than two thousand samples each class. The imbalance distribution can bring hidden danger to the recognition system, because classes with few samples can barely be recognized in the real application. The rest of the classes, about 56%, have samples distributed from 10 to 2000. Typical samples of each grades, as demonstrated in Fig. 1, are in line with common sense, for example, characters like ‘我’ and ‘有’ are popular used in daily life while ‘喟’ and ‘鸩’ are rarely used.

The shape of the text line image, especially the width size, plays an important role in recognition system. Therefore, we present the sample distribution (at logarithmic axis) with respect to image text width in Fig. 2, and draw scatter distribution of text line images with respect to their height and width in Fig. 3. In Fig. 2, we observe that images with width between 1,200 and 1,400 pixels occupies the vast majority (about 70%) of samples, while most other intervals have approximate two thousand samples. Besides, for each width interval, we visualize the character number proportion for text lines. Not surprisingly, wider images tend to possess more characters, but there is still a considerable part of wide text line images have fewer than 10 characters. In Fig. 3, part of the text line images are represented as points in the

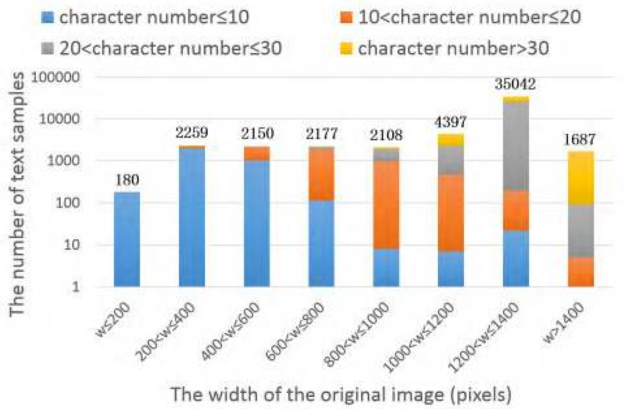

FIGURE 2. Sample distribution (at logarithmic axis) of text line image with respect to the width interval.

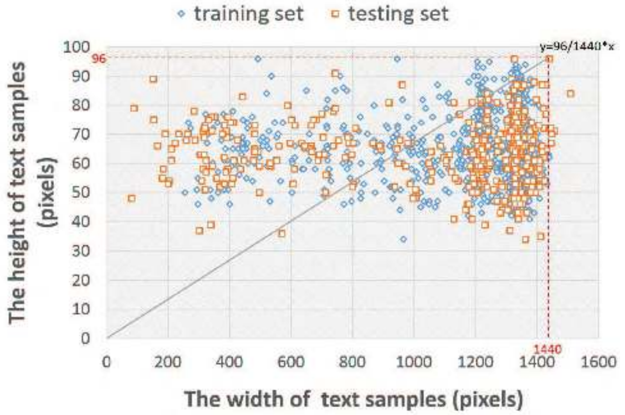

FIGURE 3. Scatter distribution of text line images with respect to their height and width.

picture with respect to their height and width. In line with the statistics in Fig. 2, the majority of the sample points in Fig. 3 have a width distribution between 1,200 and 1,400 pixels, with height ranging from 30 to 100 pixels, leaving the remaining points sparsely spread in the picture. Note that we distinguish sample points of training set from those of testing set by using different shape and color of points in Fig. 3. It can be observed that the training set and testing set of SCUT-EPT share similar sample distribution.

B. DATASET CHALLENGES

As for traditional datasets, e.g. CASIA-OLHWDB 2.0-2.2 and CASIA-HWDB2.0-2.2 [6], there are common problems such as large-scale character set, handwritten styles diversity [17] and text line mis-segmentation. When dealing with the SCUT-EPT dataset, we not only face the above-mentioned difficulties, but also have to overcome the following challenges: character erasure, text line supplement, character/phrase switching, noised background, nonuniform word size, diverse text length, which will be detailed in this section.

1) CHARACTER ERASURE

Typical examples of character erasure, also known as crossing-outs [35]–[37], can be referred to text lines (a), (b), (c), (g), (h) and (j) in Fig. 4, where we denote the erasure degree of text lines (a), (h) and (j) as *hard erasure* and the remainder as *soft erasure*. Character erasure is an inevitable problem in examination papers, therefore, it is important to let the recognition system figure out what has been modified. However, as shown in Fig. 4, text lines with *soft erasure* are very similar to the original, especially text line (c) with a simple '×' symbol in the upper right corner of the wrong characters. Since *soft erasure* can barely be distinguished from normal written characters, it can easily lead to extra prediction and *insert* error.

2) TEXT LINE SUPPLEMENT

Typical examples of text line supplement can be referred to text lines (g), (h) and (j) in Fig. 4. Text line supplement is another widespread problem in examination paper and often accompanies with character erasure problem. The additional characters usually appear right above or below the erased characters, e.g. text lines (h) and (j). Sometimes, the supplementary characters are added to the normal written sentence with special symbols, like '⋁' or '⋀', indicating the operation of text line supplement, such as text lines (h) and (j). Unfortunately, to the best of our knowledge, existing methods for offline HCTR can only handle single-line text recognition. The participation of the attention-based methods is expected to solve this kind of problem in HCTR and will be discussed in Sec. V-C.

3) CHARACTER/PHRASE SWITCHING

Typical examples of character/phrase switching can be referred to text lines (d), (e) and (f) in Fig. 4. Character/phrase switching frequently occurs in examination papers with a specific switching symbol as shown in Fig. 4. This kind of problem can hardly be resolved even with state-of-the-art technique [1], [14], [24], because it not only require the system to recognize the character, but also semantically understand the meaning of the specific switching symbol and rectify the recognition result by switching the order of the predicted characters or phrases.

4) NOISED BACKGROUND

Typical examples of noised background can be referred to text lines from (c) to (k) in Fig. 4. In the context of examination, typical background includes underlines below the characters, such as (g), (h), (j) and (k), dense grids that separate each characters, such as (c), (d), (e), (f) and (i), and printed text, such as (g). The noised background certainly brings obstacle to the recognition process, especially the printed text problem that requires the recognition system to distinguish from handwritten text. However, after investigation in our experiments, we discover that this problem is not as difficult as it seems when sufficient training samples are provided.

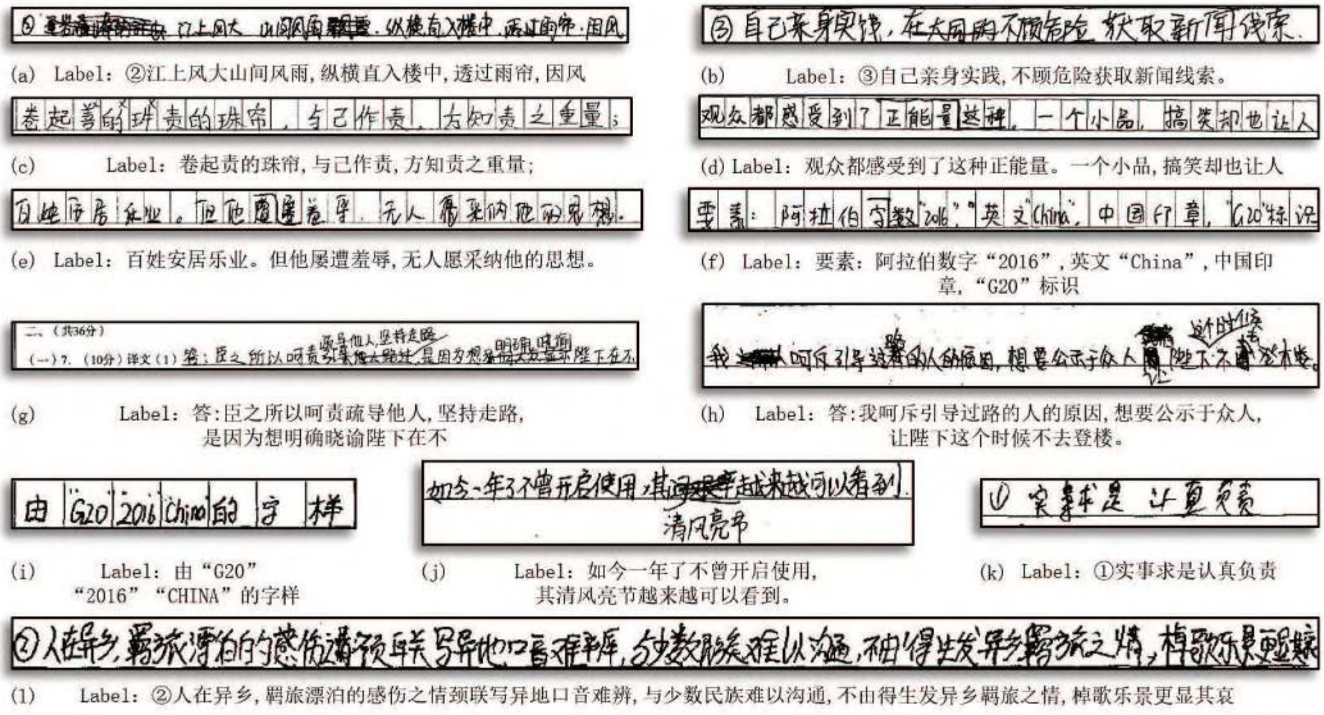
(a) Label：②江上风大山间风雨，纵横直入楼中，透过雨帘，因风
(b) Label：③自己亲身实践，不顾危险获取新闻线索。
(c) Label：卷起责的珠帘，与己作责，方知责之重量；
(d) Label：观众都感受到了这种正能量。一个小品，搞笑却也让人
(e) Label：百姓安居乐业。但他屡遭羞辱，无人愿采纳他的思想。
(f) Label：要素：阿拉伯数字“2016”，英文“China”，中国印章，“G20”标识
(g) Label：答：臣之所以呵责疏导他人，坚持走路，是因为想明确晓谕陛下在不
(h) Label：答：我呵斥引导过路的人的原因，想要公示于众人，让陛下这个时候不去登楼。
(i) Label：由“G20”“2016”“CHINA”的字样
(j) Label：如今一年了不曾开启使用，其清风亮节越来越可以看到。
(k) Label：①实事求是认真负责
(l) Label：②人在异乡，羁旅漂泊的感伤之情颈联写异地口音难辨，与少数民族难以沟通，不由得生发异乡羁旅之情，棹歌乐景更显其哀

FIGURE 4. Visualization of typical challenges in SCUT-EPT dataset, including character erasure (a, b, c, g, h, j), text line supplement (g, h, j), character/phrase switching (d, e, f), noised background (c, d, e, f, g, h, i, j, k), nonuniform word size (f, i) and unbalanced text length (i, k, l).

5) NONUNIFORM WORD SIZE

Typical examples of nonuniform word size can be referred to text lines (f) and (i) in Fig. 4. When observing text line samples of dataset SCUT-EPT, we discover that Chinese characters have relatively larger character size and character spacing size than those of punctuation, number, and English letter, which we refer to as problem of nonuniform word size. The nonuniform word size problem is very challenging even using state-of-the-art technology [1], [14], [24]. For example, within the popular CRNN [1] framework, if we allow the network to pick up the crowed and small characters, as shown in text line (f) or (i), then the stride size of fully convolutional network will be very short. However, shorter stride size will inevitably lead to more time steps of the RNN, resulting in longer training time and probably poorer performance of the network.

6) UNBALANCED TEXT LENGTH

Typical examples of unbalanced text length can be referred to text lines (i), (k) and (l) in Fig. 4. Unlike other datasets whose text line images are distributed around a certain length, the proposed SCUT-EPT dataset naturally has unbalanced text length ranging from 5 to 60 characters, as illustrated by the comparison between text line (i), (k) and (l). This is because, in examination paper, different types of questions correspond to answers of different length. The unbalanced text length problem is very unfriendly during training process, because mini-batch training strategy requires all the training samples length in a mini-batch to be exactly the same.

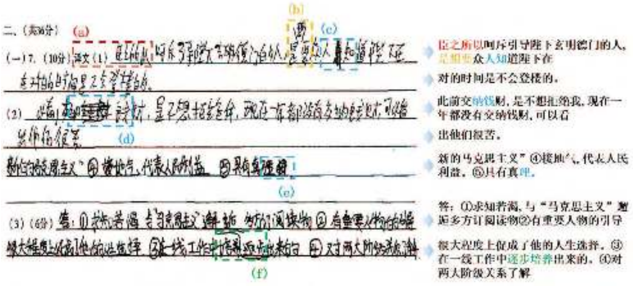

FIGURE 5. Examples of the annotation.

C. ANNOTATION METHODS

Given the above-mentioned challenges, the annotation information is expected to provide the corresponding auxiliary information to facilitate the text recognition system. Therefore, during the annotation procedure, we label the text line image with respect to the common reading habits, i.e. the annotation result is not as straightforward as characters distribute from left to right, but also considering on the special symbols from the writers. In Fig. 5, we demonstrate some typical annotation scenarios which illustrate how we perform annotation in examination paper. As shown by bounding box (a) in Fig. 5, when the text line image contains printed character such as ‘译文 (1)’, we simply neglect them in the annotation, with the hoping that our recognition system trained with these samples can distinguish handwritten text from printed text. Besides, as shown by bounding

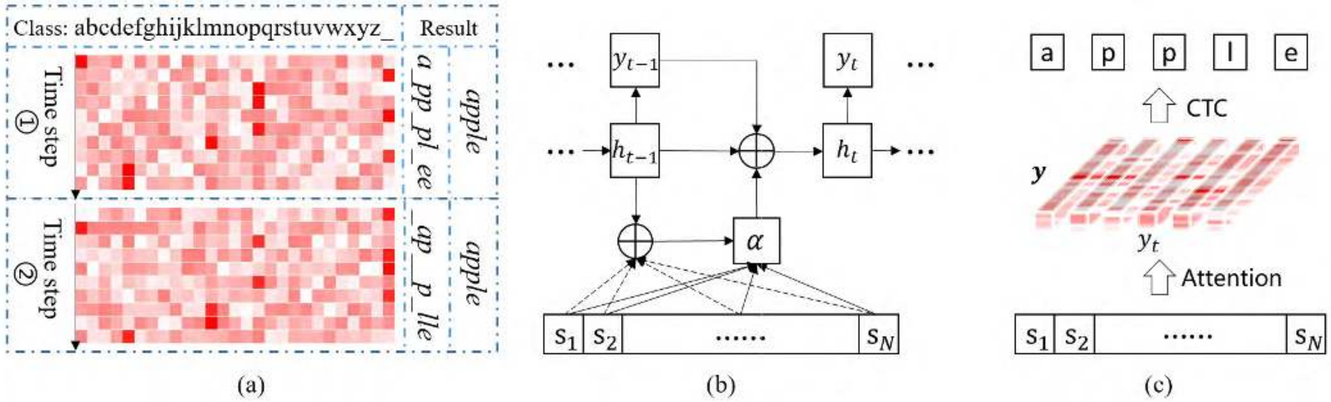

FIGURE 6. **Three kinds of transcription methods for text recognition system, including CTC decoder (a), attention mechanism (b), cascaded attention-CTC decoder (c). In figure (a), two typical examples are shown to illustrate the idea of sequence-to-sequence operation $\mathscr{B}$. In figure (b), we demonstrate the schematic of the attention mechanism. In figure (c), cascaded attention-CTC decoder is the combination of attention decoder and CTC decoder.**

box (b) with text line supplement problem, we should recognize the ‘√’ symbol and insert the additional character ‘想’ right between character ‘是’ and ‘要’. Furthermore, for character/phrase switching problem, we should follow the actual meaning of the text, and switch the corresponding characters as shown by bounding box (f). Finally, for the widespread character erasure problem, our annotation result will certainly not include them, as illustrated by bounding boxes (c), (d) and (e) in Fig. 5. Note that, except for these specific situation, we annotate the text line image exactly according to what the writer has written, completely ignoring character misspellings and grammar problem.

IV. SEQUENCE TRANSCRIPTION

In response to the above-mentioned challenges, we select the state-of-the-art text recognition framework CRNN [1] as baseline to construct our text recognition system. The proposed text recognition system consists of three parts, from bottom to top, including fully convolutional network (FCN) [39], multi-layered residual LSTM [14] and transcription layer.

Fully convolutional network can not only play the role as high-level informative feature extractor, but also take input images of arbitrary size and produce corresponding length feature sequence. Besides, FCN possesses the capability of fast inference and back-propagation by sharing convolutional feature maps layer-by-layer. Inspired by the recently proposed MC-FCRN [14] system, we apply residual LSTM for learning complex and long-term temporal information from the output feature sequence of FCN. Residual LSTM has the advantage of easily transporting gradient information in the early training stage and capturing the essential contextual information from feature sequence while not adding extra parameters nor computation burden.

For a traditional CRNN, the transcription layer generally uses CTC decoder [28], [40] to directly perform end-to-end sequential training without explicit alignment between input images and their corresponding label sequences. To evaluate the function of attention mechanism on HCTR, we further use attention decoder [29] and cascaded attention-CTC decoder [32], [33] to replace CTC as the transcription layer. Therefore, in this part, we will detail the knowledge about CTC, attention mechanism, and cascaded attention-CTC decoder.

A. CONNECTIONIST TEMPORAL CLASSIFICATION (CTC)

Connectionist temporal classification (CTC), which needs neither explicit segmentation information nor prior alignment between text line image and its text label sequence, can perform seq-to-seq transcription. After network inference, we have sequential prediction $\boldsymbol{v} = (v_1, v_2, \cdots, v_N)$ of length N for all the characters $C' = C \cup \{blank\}$, where C represents all the characters used in this problem and "blank" represents the null emission. Based on the prediction, alignments $\boldsymbol{\pi}$ is constructed by assigning a label to each time step and concatenating the labels to form a label sequence. Formally, the probability of alignments is given by

$$p(\boldsymbol{\pi}|\boldsymbol{v}) = \prod_{n=1}^{N} p(\pi_n, n|\boldsymbol{v}). \tag{1}$$

Sequence-to-sequence operation $\mathscr{B}$ first removes the repeated labels and then removes the blanks to map alignments to a transcription $\boldsymbol{l}$. Fig. 6 (a) shows two simple examples: "*a_pp_pl_ee*" and "*_ap__p_lle*", where "_" stands for "blank". In the decoding process, we first remove the adjacent repeated characters to get "*a_p_pl_e*" and "*_ap_p_le*", and then delete "_" to obtain the final results both as "*apple*". Formally, the total probability of a transcription can be calculated by summing the probabilities of all alignments that correspond to it:

$$p(\boldsymbol{l}|\boldsymbol{v}) = \sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} p(\boldsymbol{\pi}|\boldsymbol{v}). \tag{2}$$

Detailed forward-backward algorithm to efficiently calculate the probability in Eq. (2) was proposed by Graves [28], [40].

B. ATTENTION MECHANISM

Unlike CTC that can only perform sequential transcription from left to right, attention mechanism, popular in machine translation [29], is able to perform unfixed order prediction. For example, in the task of machine translation, the Chinese sequence “我昨天看了电视” can be translated into English sequence “I watched TV yesterday”, in which “昨天” corresponds to “yesterday” but they have different positions in the sentences. Attention mechanism works in line with the way we perceive things, and has recently exhibited outstanding performance in the fields of speech recognition [41], scene text recognition [30], [31], image processing [42], etc.

As shown in Fig. 6 (b), we assume that the CRNN output sequence (annotation vectors) is $\boldsymbol{s} = (s_1, s_2, \ldots, s_N)$, the previous hidden state and output are h_{t-1} and y_{t-1}, respectively. Then, at time step t, the attention score α_t is calculated first as:

$$e_{t,j} = V_a \phi(W_a s_j + U_a h_{t-1}) \tag{3}$$

$$\alpha_{t,j} = \frac{exp(e_{t,j})}{\sum_{k=1}^{W} exp(e_{t,k})} \tag{4}$$

where ϕ represents the hyperbolic function, V_a, W_a and U_a are trainable parameters, and $j = 1, \cdots, N$. Next, we can get the context vector c_t by calculating the weighted average of annotation vectors:

$$c_t = \sum_{j=1}^{N} \alpha_{t,j} s_j \tag{5}$$

Afterward, the recurrent neural network (GRU/LSTM) will together consider the context vectors c_t, previous hidden state h_{t-1}, and previous prediction y_{t-1} to compute the t-th hidden state h_t and its prediction y_t as follows:

$$h_t = \sigma(W_o E(y_{t-1}) + U_o h_{t-1} + C_o c_t) \tag{6}$$

$$y_t = \text{Generate}(h_t) \tag{7}$$

where Generate represents a feed-forward network, σ represents the sigmoid function, W_o, U_o and C_o are trainable parameters, and E is a character-level embedding matrix to embed the previous predicted character.

C. CASCADED ATTENTION-CTC DECODER

As illustrated in [32]–[34], CTC system relies only on the hidden feature vector at the current time step to make predictions, i.e., the output predictions are independent given the input feature sequence. In their work, they combine attention mechanism and CTC network to alleviate this drawback in the application such as lipreading recognition and speech recognition.

As shown in Fig. 6 (c), cascaded attention-CTC decoder first applies attention mechanism to align the annotation vectors $\boldsymbol{s} = (s_1, s_2, ..., s_N)$ to the context vectors $\boldsymbol{c} = (c_1, c_2, ..., c_T)$. Based on the context vector $\boldsymbol{c}$, we calculate the hidden state $\boldsymbol{h}$ and its prediction $\boldsymbol{y}$. After that, we can update the formulation to calculate the probability of alignment and transcription as follows:

$$p(\boldsymbol{\pi}|\mathbf{y}) = \prod_{t=1}^{T} p(\pi_t, t|\mathbf{y}) \tag{8}$$

$$p(\boldsymbol{l}|\mathbf{y}) = \sum_{\boldsymbol{\pi}:\mathscr{B}(\boldsymbol{\pi})=\boldsymbol{l}} p(\boldsymbol{\pi}|\mathbf{y}). \tag{9}$$

Therefore, the training is achieved by minimizes the negative penalized log-likelihood:

$$L_{atten-ctc} = -\sum_{(\boldsymbol{x},\boldsymbol{l})\subset Q} \ln p(\boldsymbol{l}|\mathbf{y}), \tag{10}$$

where Q represents the training set.

V. EXPERIMENTS

A. EXPERIMENTAL SETTING

Considering the complexity of the HCTR problem, we do not use the original CRNN directly, but construct our own framework with customized FCN, multi-layered residual LSTM, and transcription layer. Specifically, our baseline network has the following network architecture:

$32C3 - MP2 - 64C3 - MP2 - 128C3 - MP2 - 128C3 - 256C3 - 512C3 - MP2 - 512C3 - 512C(3*1) - 512C(2*1) - ResidualLSTM * 3 - IP7358 - CTC$,

where xCy represents a convolutional layer with kernel size of $y * y$ and output number of x, MPx denotes a maximum pooling layer with kernel size of x, IPx means a prediction layer (fully connected layer) with output number of x, and so on. In particular, the prediction layer has 7,358 kernels, of which 7,356 kernels correspond to the character set [12], one represents the blank symbol for CTC, and the last one indicates all outlier characters. The reason why we use 7,356 classes instead of 4,250 classes in Table 1 is that our synthetic data covers the entire 7,356 class character set. In this section, we design extensive experiments to analyze the effect of different factors, including image resizing methods, whether to use synthetic data, different output feature length, and effect of fully connected layers. Furthermore, to evaluate the effect of transcription layer on HCTR, we conduct experiments to compare CTC, attention mechanism and cascaded attention-CTC decoder.

In this part, we briefly introduce the above-mentioned factors. During training process, we preprocess the images and resize them all to 96 ∗ 1440, which corresponds to the intersection of the dash red lines in Fig. 3. Specifically, we compare two image resizing methods. The first method (denoted as “$R1$”) places the images in the center without distortion and fills them with white background to construct text line images with the shape of 96 ∗ 1440. The second method (denoted as “$R2$”) is the same with the first method except that the images are placed randomly inside the 96 ∗ 1440-shape text line images. However, if the original image is larger than shape of 96 ∗ 1440, both of the methods simply reshape the image to shape of 96∗1440. Furthermore, by changing the kernel size of first three pooling layers,

TABLE 2. Comparison among various attributes, including different image resizing methods (Resize) "*R*1" and "*R*2", and whether to enrich training set with synthetic data (Enrich). "Iterations" represents the number of iterations required for the network to reach convergence.

Experiments	Resize	Enrich	CR(%)	AR(%)	Iterations
baseline	*R1*	×	77.92	74.50	120,000
(a)	*R2*	×	78.60	75.37	120,000
(b)	*R1*	√	79.03	74.90	200,000
(c)	*R2*	√	**80.26**	**75.97**	200,000

we can get prediction of different sequence length. Finally, since only one fully connected layer is used as the final prediction layer, we try to use more fully connected layers and compare their effects. For the attention mechanism, we use the same implementation details as the baseline CRNN-based system, except that the CTC decoder is replaced with attention decoder or cascaded attention-CTC decoder.

In our experiment, we also use isolated characters from CASIA-HWDB1.0-1.2 [6] to synthesize the semantic-free text dataset with 188,014 text line images. During the synthesizing stage, each time, a character sample was selected from dataset CASIA-HWDB1.0-1.2 [6] and placed next to previous characters with their centroids aligned approximately in a straight line. For some special symbols, like comma and period, we placed them to the bottom right position of previous character.

For all our experiments, we do not use language model. Besides, we use the correct rate (CR) and accuracy rate (AR) proposed by ICDAR2013 competition [12] as network recognition performance criterion. They are given by:

$$CR = (N - D_e - S_e)/N \tag{11}$$

$$AR = (N - D_e - S_e - I_e)/N \tag{12}$$

where N is the total number of characters in the ground-truth text lines, D_e, S_e, and I_e represent deletion errors, substitution errors and insertion errors, respectively. Additionally, most experiments require approximately one day to achieve convergence based on the GeForce Titan-X GPU and PyTorch [43] deep learning framework.

B. NETWORK ARCHITECTURE AND DATA AUGMENTATION

1) EFFECT OF IMAGE RESIZING METHOD AND SYNTHETIC DATA

Table 2 presents some experiment results that compare among various attributes, including different image resizing methods (denoted as "Resize"), and whether to enrich training set with synthetic data (denoted as "Enrich"). By comparing experiment (a) and baseline, we can see that image-resizing method "*R*2" shows superior performance over "*R*1". This is because randomly adding white background around the image can make the network less sensitive to the position of characters when recognizing, thereby improving the generalization of the network. Comparison between experiment (b) and baseline shows that our recognition network can still benefit from the additional training samples, even though the synthesized text line images are quite different from those of examination papers with noised background. This is because CASIA-HWDB1.0-1.2 covers the character set with samples distributed evenly over each classes. However, the network needs more iterations to reach convergence when using synthetic dataset. Experiment (c) with image resizing method "*R*2" and enriched training set provides the best result for the proposed dataset SCUT-EPT.

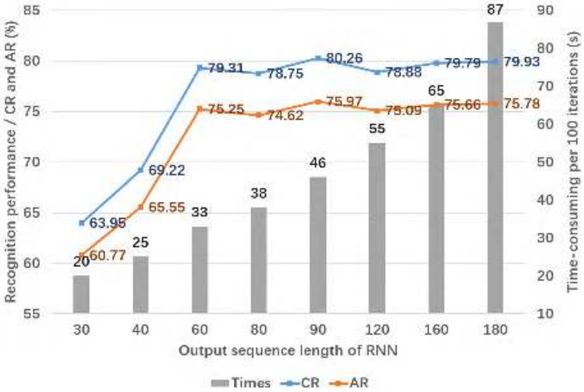

FIGURE 7. Recognition performance (correct rate and accuracy rate) and time consumption with respect to the output sequence length of RNN.

It is noteworthy that the recognition network used in the paper can achieve state-of-the-art result (CR of 92.25% and AR of 91.76% without language model) in the popular ICDAR2013 competition set [12]. This in return verifies the challenge and significance of the proposed dataset SCUT-EPT.

2) EFFECT OF DIFFERENT FEATURE SEQUENCE LENGTH

In this section, we conduct experiments to compare the effects of different feature sequence length. As demonstrated in Fig. 7, the recognition performance of the system is very poor when the sequence length is short, e.g., sequence length of 30 with AR of 60.77. However, the recognition performance improves very fast from sequence length of 30 to 60 and becomes stable after that. This is because the character number in text line images of the proposed SCUT-EPT is around 25 (see Fig. 2), when sequence length is too short, deletion errors can easily occur for text line images with long sequence length.

As shown Fig. 7, the time consumption of network training gradually increases as the sequence length becomes longer. Sequence length of 90 strikes a balance between time consumption and recognition performance, so it is used as default setting in the following experiments.

3) EFFECT OF FULLY CONNECTED LAYERS:

In this section, we evaluate the role of the fully connected layers, which have kernel size of 512 and are placed between multi-layered residual LSTM and the final prediction layer. As shown in Table 3, it is surprising to see that network performance decline gradually as the number of fully connected layers increases. Considering the memory size, training time consumption, and most importantly, recognition

TABLE 3. Comparison of fully connected layer number. "Size" represents the model size of each network and "Iterations" denotes the number of iterations required for the network to reach convergence.

Number of FC Layers	CR(%)	AR(%)	Size (M)	Iterations
None	**80.26**	**75.97**	**62**	**20**
1	79.36	75.34	63	24
2	78.28	74.59	64	27
3	77.72	73.49	65	33

TABLE 4. Comparison among different transcription methods, where "Att-time" denote the total time steps during attention process, and "Enrich" represents whether to enrich training set with synthetic data.

Transcription Method	Att-time	Enrich	CR(%)	AR(%)
CTC	-	×	78.60	75.37
		√	**80.26**	**75.97**
Attention mechanism	adapted	×	69.83	64.78
		√	73.10	67.04
Cascaded attention-CTC	90	×	35.14	33.03
		√	not converge	
	30	×	54.09	48.98
		√	60.20	55.64

performance, we suggest not to use fully connected layer between multi-layered residual LSTM and the final prediction layer for HCTR problem.

C. TRANSCRIPTION METHODS

In Table 4, we provide comprehensive investigations on the popular transcription methods, including CTC, attention mechanism and cascaded attention-CTC decoder.

1) ATTENTION MECHANISM

In English scene text recognition problem [30], [31], attention-based methods exhibit superior performance than CTC-based methods. However, in Table 4, it is observed that attention mechanism has much worse performance than that of CTC. For example, without enriching training set, network trained with attention mechanism has a CR of 69.83%, much worse than CTC-based network with a CR of 78.60%.

There are two main differences between English scene text recognition and HCTR problems: class number and feature sequence length. For scene text recognition problem [30], [31], there are only 36 classes, including 10 numbers and 26 letters. However, there are often thousands of classes in handwritten Chinese text recognition problem. Furthermore, scene text recognition only requires network to make word-level prediction, while HCTR makes prediction at text-level; thus, these two problems have quite different prediction sequence length. The latter difference is relatively more important, because attention inherently has the drawback that it requires the prediction exactly the same with ground-truth. For example, if we predict the text line image with ground-truth 'computer' as 'comuter', then we should have only one deletion. However, attention mechanism will only consider 'com' as the right prediction, while the remainder 'uter' is wrong prediction, because the remainder prediction 'uter' is not the same with the ground-truth 'puter' at each position.

TABLE 5. Comparison of previous methods for text recognition problem.

Methods	CR(%)	AR(%)
Attention [44], [45]	73.10	67.04
CNN+CTC [46]	75.46	70.81
CNN+MDLSTM+CTC [25]	78.30	73.26
CNN+MDirLSTM+CTC [47]	78.53	73.65
CNN+LSTM+CTC [1], [14], [24]	**80.26**	**75.97**

2) CASCADED ATTENTION-CTC DECODER

For cascaded attention-CTC, we first set attention times to 90, but the network shows poor performance, or even cannot converge, as shown in Table 4. However, when we decrease attention times to 30, its performance becomes much better, but still not as good as attention mechanism and CTC decoder. Note that when we decrease the attention times, it will inevitably bring deletion errors to those text line samples with more than 30 characters. In the other hand, when we increase the attention times, it will decrease the accuracy rate most of the time, and sometimes cause the network not to converge.

The combination of attention mechanism and CTC is a novel idea in speech-related field [32]–[34]. Specifically, Das *et al.* [33] attributes the inferior performance of the individual CTC decoder to its conditional independence prediction assumption. However, this problem is relatively less important in HCTR problem. Actually, we perform an additional experiment using only FCN and CTC transcription layer, resulting in a CR of 75.46% and an AR of 70.81%, as shown in Table 5. In other word, multi-layered residual LSTM improves network performance from CR of 75.46% to CR of 80.26%. Therefore, we consider that multi-layered residual LSTM has already learned context information from previous time steps to benefit current time step prediction. This may be the reason why cascaded attention-CTC decoder does not work well in HCTR problem.

D. COMPARISON WITH PREVIOUS METHODS

To further reveal the challenge of the SCUT-EPT dataset, we reproduce state-of-the-art seq-to-seq methods for text recognition problem on SCUT-EPT in Table 5. Since our solution for SCUT-EPT dataset is based on deep-learning technique and deep-learning-based methods dominate state-of-the-art result on most of the handwritten datasets, we only make comparison for this kind of methods in this section.

As shown in Table 5, although attention-based methods demonstrate state-of-the-art result for western text recognition [44], [45], it exhibits relatively poor result for the HCTR problem, as compared to CTC-based methods. This is because missing or superfluous characters can easily cause misalignment problem and mislead the training process for attention module [31]. This phenomenon becomes more severe in HCTR problem, in which Chinese text length is much longer (compared to western word recognition) and the character set is much larger. Further, we can also observe that pure CNN architecture with CTC cannot make full use of context information without the assistance of the recurrent neural

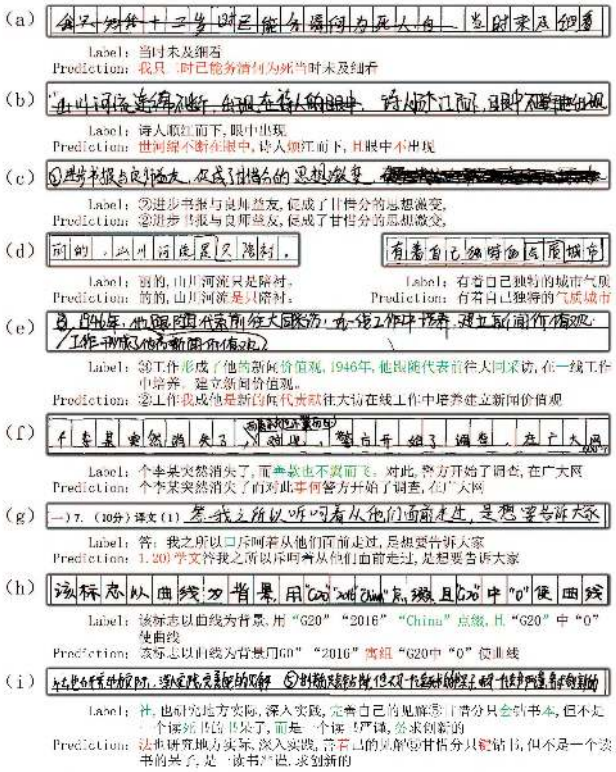
(a)
Label: 当时未及细看
Prediction: 我只 时已能务清何为死当时未及细看
(b)
Label: 诗人顺江而下，眼中出现
Prediction: 世河绵不断在眼中，诗人倾江而下，且眼中不出现
(c)
Label: ②进步书报与良师益友，促成了甘惜分的思想激变，
Prediction: ②进步书报与良师益友，促成了甘惜分的思想激变，
(d)
Label: 前的，山川河流只是陪衬。
Prediction: 前的，山川河流是只陪衬。
Label: 有着自己独特的城市气质
Prediction: 有若自己独特的气质城市
(e)
Label: ③工作形成了他的新闻价值观，1946年，他跟随代表前往大同采访，在一线工作中培养、建立新闻价值观。
Prediction: ③工作我成他是新的闻代贡献往大访在线工作中培养建立新闻价值观
(f)
Label: 个李某突然消失了，而善款也不翼而飞。对此，警方开始了调查，在广大网
Prediction: 个李某突然消失了而对此李何警方开始了调查，在广大网
(g)
一）7. （10分）译文（1）
Label: 答：我之所以口斥呵着从他们面前走过，是想要告诉大家
Prediction: 1. 20) 学文答我之所以斥呵着从他们面前走过，是想要告诉大家
(h)
Label: 该标志以曲线为背景，用“G20” “2016” “China”点缀，且“G20”中“0”使曲线
Prediction: 该标志以曲线为背景用G0” “2016” 离组 “G20中 “0” 使曲线
(i)
Label: 社，也研究地方实际，深入实践，完善自己的见解⑤甘惜分只会钻书本，但不是一个读死书的书呆子，而是一个读书严谨，发求创新的
Prediction: 法也研究地方实际，深入实践，苦若自的见解⑤甘惜分只能钻书，但不是一个读书的呆子，是一读书严谨，求到新的

FIGURE 8. Visualization of recognition results.

network, thereby show inferior results to methods equipped with MDLSTM [25], MDirLSTM [47] or LSTM. Both MDLSTM and MDirLSTM model possess advantage of two-dimensional context learning and share similar performance on the SCUT-EPT dataset. Lastly, we observe that LSTM-based seq-to-seq model shows better performance than that of MDirLSTM-based model. This is probably because MDirLSTM was initially designed for western language word recognition. Two-dimensional spatial context learning based on MDirLSTM is necessary for high performance of western language written in a cursive and overlapping manner, which, however, is not very critical for HCTR problem.

E. RESULTS ANALYSIS

In Fig. 8, we investigate some recognition result samples to gain additional insights, where green color indicates *deletion* error and red color indicates *substitution* and *insertion* error. The challenges discussed in Sec. III-B are the main causes of the error predicted results.

By comparing examples (a), (b) with (c), we can observe that softer character erasure will bring more *insertion* errors, because hard erasure is easier to be captured by the recognition system and can avoid being recognized. Next, for noised background problem, recognition can correctly distinguish characters from background, like underlines and grids, but fail to filter out a few printed samples as shown in example (g). Character erasure and noised background share the same problem that needs the network 'observe' very carefully to distinguish the normal characters from erasure characters, printed text or background. Therefore, feature extraction networks like ResNet [48] and DenseNet [49] may be good alternatives of FCN to our text recognition network.

Error examples (d) are caused by character/phrase switching problem and can barely be rectified, as their switching symbol is easily ignored by recognition model. Next, supplemental problem is very common in examination paper, and recognition system can hardly provide tolerable prediction results for examples like (e) and (f). The additional text is usually ignored, or even worse, preventing the parallel text from being recognized. Character/phrase switching and character supplemental share the same problem that requires the system not to simply recognize text line images from left to right, but also in more complex spatial order with respect to the specific symbol. To the best of our knowledge, attention mechanism, which can naturally decode the text image in arbitrary order, should have great potential to solve this problem theoretically. However, as described in Sec V-C, neither attention decoder nor cascaded attention-CTC decoder can achieve acceptable performance on the dataset SCUT-EPT. Therefore, we think this problem is the most challenging one that reveals the limitation of existing advanced text recognition technology and deserves further research.

Examples like (h) and (i) in Fig. 8 suffer from nonuniform word size problem. There are characters occasionally missing in prediction results, especially those small and dense. This problem can be alleviated by extending the feature sequence length, but at the cost of slower convergence and training speed as shown in Fig. 7. Other alternatives may allow the system to adaptively choose recognition modules of different receptive fields with respect to actual character sizes of text line images.

VI. CONCLUSION

In this paper, we present a new dataset SCUT-EPT for examination papers, covering numerous novel challenges nonexistent in ordinary HCTR datasets, including character erasure, text line supplement, character/phrase switching, noised background, nonuniform word size, unbalanced text length. In the body of the paper, we not only provide diagrams to analyze the dataset SCUT-EPT, but also discuss the above-mentioned difficulties in detail with sample visualization. In the experiments, we investigate dataset SCUT-EPT with our text recognition system customized from the popular CRNN, but only observe poor performance, which verifies the challenge and significance of the SCUT-EPT dataset. Besides, we provide a comprehensive investigation on three popular transcription methods on HCTR problem, including CTC, attention mechanism, and cascaded attention-CTC decoder. However, we discover that the attention-based decoding methods perform poorly in HCTR with large-scale character set; thus, how to design an effective attention decoding model for HCTR is still an open problem. Furthermore, we provide visualization of typical text line images and their recognition results, with briefly discussion on the cause of the errors and constructive suggestions for the problems.

We hope that the dataset SCUT-EPT brings new challenge to the community and promotes the research progress. In future, we will focus on solving the challenges in this dataset, especially for text line supplement problem which reveals the single-line recognition limitation of existing technology and deserves further exploration.

ACKNOWLEDGMENT

(Yuanzhi Zhu and Zecheng Xie contributed equally to this work.)

REFERENCES

[1] 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, Nov. 2017.

[2] E. Grosicki, M. Carre, J.-M. Brodin, and E. Geoffrois, "Rimes evaluation campaign for handwritten mail processing," in *Proc. 11th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, 2008, pp. 1–6.

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

[4] M. Pechwitz and V. Margner, "Baseline estimation for arabic handwritten words," in *Proc. 8th Int. Workshop Frontiers Handwriting Recognit.*, Aug. 2002, pp. 479–484.

[5] S. A. Mahmoud *et al.*, "KHATT: Arabic offline handwritten text database," in *Proc. Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Sep. 2012, pp. 449–454.

[6] 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. (ICDAR)*, Sep. 2011, pp. 37–41.

[7] T. Su, T. Zhang, and D. Guan, "HIT-MW dataset for offline Chinese handwritten text recognition," in *Proc. Int. Workshop Frontiers Handwriting Recognit.*, Oct. 2006, pp. 1–5.

[8] T. Matsushita and M. Nakagawa, "A database of on-line handwritten mixed objects named 'Kondate,'" in *Proc. 14th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Sep. 2014, pp. 369–374.

[9] M. Nakagawa and K. Matsumoto, "Collection of on-line handwritten japanese character pattern databases and their analyses," *Document Anal. Recognit.*, vol. 7, no. 1, pp. 69–81, 2004.

[10] M. Liwicki and H. Bunke, "IAM-OnDB—An on-line English sentence database acquired from handwritten text on a whiteboard," in *Proc. 8th Int. Conf. Document Anal. Recognit.*, Aug./Sep. 2005, pp. 956–961.

[11] 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, no. 1, pp. 53–64, 2011.

[12] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "ICDAR 2013 Chinese handwriting recognition competition," in *Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2013, pp. 1464–1470.

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

[14] 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, pp. 1903–1917, Aug. 2018.

[15] H. Yang, L. Jin, W. Huang, Z. Yang, S. Lai, and J. Sun, "Dense and tight detection of Chinese characters in historical documents: Datasets and a recognition guided detector," *IEEE Access*, vol. 6, pp. 30174–30183, 2018.

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

[17] R. Dai, C. Liu, and B. Xiao, "Chinese character recognition: History, status and prospects," *Frontiers Comput. Sci.*, vol. 1, no. 2, pp. 126–136, 2007.

[18] X. Ren, Y. Zhou, Z. Huang, J. Sun, X. Yang, and K. Chen, "A novel text structure feature extractor for Chinese scene text detection and recognition," *IEEE Access*, vol. 5, pp. 3193–3204, 2017.

[19] 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.

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

[21] X.-D. Zhou, D.-H. Wang, F. Tian, C.-L. 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, Oct. 2013.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] 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 Recognit.*, Jun. 2016, pp. 4168–4176.

[24] Z. Xie, Z. Sun, L. Jin, Z. Feng, and S. Zhang, "Fully convolutional recurrent network for handwritten Chinese text recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 4011–4016.

[25] R. Messina and J. Louradour, "Segmentation-free handwritten Chinese text recognition with LSTM-RNN," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 171–175.

[26] G. Zhu, "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[27] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017.

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

[29] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: https://arxiv.org/abs/1409.0473

[30] 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 Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5086–5094.

[31] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou. (2018). "Edit probability for scene text recognition." [Online]. Available: https://arxiv.org/abs/1805.03384

[32] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 548–555.

[33] A. Das, J. Li, R. Zhao, and Y. Gong. (2018). "Advancing connectionist temporal classification with attention modeling." [Online]. Available: https://arxiv.org/abs/1803.05563

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

[35] B. B. Chaudhuri and C. Adak, "An approach for detecting and cleaning of struck-out handwritten text," *Pattern Recognit.*, vol. 61, pp. 282–294, Jan. 2017.

[36] N. Bhattacharya, U. Pal, and P. P. Roy, "Cleaning of online Bangla free-form handwritten text," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 1, p. 8, 2017.

[37] N. Bhattacharya, V. Frinken, U. Pal, and P. P. Roy, "Overwriting repetition and crossing-out detection in online handwritten text," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 680–684.

[38] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "AON: Towards arbitrarily-oriented text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5571–5579.

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

[40] 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, May 2009.

[41] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[42] D. Bhowmik, M. Oakes, and C. Abhayaratne, "Visual attention-based image watermarking," *IEEE Access*, vol. 4, pp. 8002–8018, 2016.

[43] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. NIPS-W*, 2017, pp. 1–4.

[44] T. Bluche, J. Louradour, and R. Messina, "Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 1050–1055.

[45] 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.*, to be published.

[46] F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu. (2017). "Scene text recognition with sliding convolutional character models." [Online]. Available: https://arxiv.org/abs/1709.01727

[47] Z. Sun, L. Jin, Z. Xie, Z. Feng, and S. Zhang, "Convolutional multi-directional recurrent network for offline handwritten text recognition," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 240–245.

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

[49] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, vol. 1, no. 2, Jun. 2017, p. 3.

YUANZHI ZHU received the B.S. degree in electronics and information engineering from the South China University of Technology, where he is currently pursuing the master's degree in electronic and communication engineering. His research interests include machine learning, document analysis and recognition, and computer vision.

ZECHENG XIE received the B.S. degree in electronics and information engineering from the South China University of Technology, in 2014, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include machine learning, document analysis and recognition, computer vision, and human-computer interaction.

LIANWEN JIN (M'98) 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 College of Electronic and Information Engineering, South China University of Technology. He has authored over 100 scientific papers. His research interests include handwriting analysis and recognition, image processing, machine learning, and intelligent systems. He is a member of the IEEE Computational Intelligence Society, the IEEE Signal Processing Society, and the IEEE Computer Society. He has received the New Century Excellent Talent Program of MOE Award and the Guangdong Pearl River Distinguished Professor Award.

XIAOXUE CHEN received the B.S. degree in electronics and information engineering from the South China University of Technology, where she is currently pursuing the master's degree in signal and information processing. Her research interests include machine learning and computer vision.

YAOXIONG HUANG received the B.S. degree in electronics and information engineering from the South China University of Technology, where he is currently pursuing master's degree in communication and information system. His research interests include machine learning and computer vision.

MING ZHANG received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2000 and 2011, respectively. He was an Engineer with Cisco Inc., from 2000 to 2008. He joined Ailibaba Group as a Senior Algorithm Expert, in 2008. From 2013 to 2016, he was with Beijing Oriental Junguan Technology Co., Ltd., as the CTO. He founded AbcPen Inc., and served as the Chief Architect, in 2017.