

Character Recognition for Tibetan Ancient Books Based on Deep Learning Approaches

Rinchen Dongrub^a, Dorje Tashi^a, Thupten Tsering^b, Nyima Tashi^{a,*}, YU Yong-bin^c,
CHENG-Jian^d

^a*School of Information Science and Technology, Tibet University, Lhasa, 850000*

^b*The state key laboratory of Tibet an intelligent information processing and application, Xining, 810008*

^c*School of Information and Software Engineering, University of Electronic Science and Technology, Chengdu, 610054*

^d*School of Information and Communication Engineering, University of Electronic Science and Technology, Chengdu, 611731*

Abstract

In this paper, a novel method for Tibetan ancient book image character recognition based on deep neural networks (CNN and RNN) was proposed. To solve the problem of ultra-long text line ancient book recognition, the ultra-long text is dynamically split into multiple sub-strings with the sliding window based Tibetan ancient book text line splitting technology, and the adjacent sub-string overlapping characters are processed based on character location information. The string recognition model of Tibetan ancient books based on ResNet[37] (Residual Network) and Bi-LSTM[38] (Bidirectional Long Short-Term Memory) recurrent neural network (RNN) was constructed to solve the problem of character recognition of ancient books with poor image quality, serious adhesion of adjacent characters, and large overlap between up and down lines. Moreover, Through experimental comparison and analysis, the macro average accuracy of our proposed method in modern Tibetan publication print recognition has reached 97.27%, which is 12.01 and 10.95 higher than TH-OCR (85.26%) and UTibet-OCR-2008 (86.32%) respectively. The average accuracy rate of the proposed method for the recognition of Tibetan ancient woodcut characters has reached 92.60%, and the average accuracy rate for the recognition of Tibetan ancient handwritten characters has reached 85.47%, which demonstrates that our proposed method has achieved an important breakthrough in the field of Tibetan ancient character recognition at home and abroad, and has been open to the public for free trial.

Keywords: Computational Linguistics; Tibetan ancient books; character recognition; Residual Network; Bi-LSTM

*Corresponding author: Nyima Tashi (Email: nmzx@utibet.edu.cn; ORCID: 0000-0001-9288-6600)

1. Introduction

China, as a multi-ethnic nation, has fostered a rich and diverse cultural heritage, with its splendid Chinese culture emerging from the collaborative creation, ongoing development, and evolution by various ethnic groups, predominantly grounded in the Han culture of the Central Plains. Ancient manuscripts represent a pivotal component of this cultural repository, embodying the collective wisdom and inventive spirit of the Chinese nation. These manuscripts serve as invaluable cultural assets, offering essential insights for the construction of a contemporary socialist culture and the shaping of new societal paradigms. The scholarly exploration into the digitization and textual analysis of ancient literary resources from minority ethnic groups is critical. Such research endeavors not only deepen the global and domestic comprehension of China's traditional ethnic cultures but also invigorate the national ethos, augment the promotion of national cultural values, and enhance the proficiency and capacity in safeguarding, excavating, and leveraging the ancient textual heritage of China's ethnic minorities.

Contemporary advancements in the field of character recognition within Chinese ancient texts, including oracle bone inscriptions and the identification of variant characters, alongside applications in digital processing of these ancient manuscripts, have significantly contributed to the foundational efforts for the preservation and enhancement of China's esteemed traditional culture. The Tibetan script, developed during the Tubo (Tibetan) Dynasty in the 7th century AD, is a phonetic writing system. Today, it remains predominantly in use within Tibetan regions of China and in several southwestern neighboring countries. This significant aggregation of Tibetan and other ethnic documents underscores the library's pivotal role in preserving the linguistic and cultural heritage of China's diverse ethnic groups, facilitating scholarly research and promoting cultural understanding and appreciation both domestically and internationally [2].

The study of text recognition for ancient Tibetan manuscripts holds a pivotal role in the digital preservation and content excavation of these historical texts. Noteworthy initiatives include projects undertaken by Tibet University among others, such as the "Tibetan Literature Resources Digitization Technology Integration and Application Demonstration Project" under the National Key Research and Development Plan. This project team has successfully digitized nearly 10,000 Tibetan texts, of which approximately 2,000 are ancient Tibetan medical manuscripts recognized as text. However, research on the recognition of varied fonts, handwritten texts, and social media content within ancient Tibetan documents remains scarce or entirely absent. Internationally, institutions such as Alcatel-Lucent Bell Labs, the University of Virginia, and the Tohoku Institute of Technology in Japan have initiated foundational studies in Tibetan text recognition, indicating a growing interest and the potential for significant advancements in this field. Google's foray into Tibetan OCR technology has been successfully integrated into its Google Books and Google Drive projects, showcasing the potential for widespread application and accessibility. Concurrently, a recognition system developed by the University of California, Berkeley, as documented in literature [3], alongside studies [4][5], delineates the comprehensive process from image acquisition to the recognition of Tibetan printed text, presenting a holistic approach to this complex task.

Domestically, scholarly works, such as literature [6], delve into the nuances of Tibetan character recognition by exploiting the unique syllabic structure of the Tibetan script. Further contributions by Wang Haojun et al. [7] include the development of a preprocess-

ing methodology specifically tailored to the glyphs and writing characteristics of printed Tibetan texts, indicating a customized approach to OCR. Additionally, literature [8][9] employs computer statistical methods to analyze Tibetan glyphs, proposing an innovative character feature selection scheme that underscores the significance of linguistic and scriptural nuances in OCR technology.

Wu Gang and colleagues have put forward a preprocessing and feature extraction scheme dedicated to printed Tibetan OCR, by thoroughly analyzing Tibetan glyphs, marking a critical step towards refining the accuracy and efficiency of recognition systems. Although literature [10][11] discusses recognition and rule-based post-processing methods tailored to Tibetan multi-font prints, there remains a notable gap in addressing the recognition of text within ancient Tibetan manuscripts. This highlights the ongoing challenge and the necessity for focused research and development efforts aimed at bridging this gap, ensuring the preservation and accessibility of Tibet's rich literary heritage through advanced OCR technologies.

The adoption and innovation in deep learning technologies have significantly propelled the domain of handwritten text detection within ancient Tibetan manuscripts. Scholars are leveraging split-based deep learning text detection technologies such as PSENet, PixelLink, and PANNet [12] to enhance the detection accuracy of handwritten texts in these ancient documents. A particular focus has been placed on the woodcut editions of Tibetan ancient books in literature [13], where the application of the CTC (Connectionist Temporal Classification) recognition algorithm in Tibetan text recognition has demonstrated promising results on test datasets.

The methodologies applied in Tibetan OCR (Optical Character Recognition) research encompass both traditional recognition techniques and modern deep learning approaches. These methods often incorporate studies on Tibetan spelling grammar, the structural and writing characteristics of Tibetan characters, and the construction rules of these characters. Notably, Nyima Tashi [14][15] has provided a detailed analysis based on the construction rules of Tibetan characters, offering theoretical insights into the formal language and application of Tibetan spelling grammar for character recognition. Similarly, Wang Daohui [16] has engaged in empirical research on Tibetan character feature extraction, exploring various methods of analysis.

Furthermore, advancements have been made in the synthesis of Tibetan ancient book layout analysis technology and training data generation through deep learning by researchers such as Yongcuo [17][18]. In addressing the challenges posed by degraded image quality of ancient manuscripts, Lu Yuqi [19] has developed a multi-size document image super-resolution reconstruction method specifically for Tibetan ancient books. Liu Yashan [20] has introduced algorithms for multi-line text recognition in these manuscripts, focusing on implicit split and no-split techniques, thus verifying the efficacy of these methods in practical scenarios.

Despite these advancements, the field of ethnic minority text information processing largely remains in the glyph processing phase, with the digitization level of ancient literary resources generally low. Notably, there is a pronounced scarcity of research dedicated to the scanning and recognition of ancient document texts. The application of Tibetan character formation rules and the intricate structure of sentences in ancient document recognition technology research is still nascent. Consequently, the recognition accuracy and generalization capabilities of current technologies do not fully meet the practical requirements for the digital preservation and utilization of Tibetan ancient documents,

indicating a significant scope for future research and development in this critical area.

With the further development of deep learning technology, in the field of text recognition, this technology has advantages over traditional recognition methods and has achieved better recognition performance. Currently, how to use deep learning to perform end-to-end learning and abandon intermediate steps based on manual rules to improve the text string recognition effect has become a research hotspot [21]. The strategy of CRNN (Convolutional Recurrent Neural Network) has gradually attracted attention. According to the advantages of each model, CNN [22] is used to extract local features of images; RNN [23] is used to deal with sequence problems in images, and research in the field of text string recognition has begun to a substantial effect. This method has been used in various fields of Tibetan text recognition [24][25][26] and has achieved good performance.

Based on this, this paper proposes a CRNN-based text recognition method for Tibetan ancient books by analyzing the layout and text characteristics of Tibetan ancient books. Its main work is summarized as follows:

I. In order to solve the problem of identifying ultra-long text lines in ancient books, the sliding window-based text line splitting technology of Tibetan ancient books is used. First, the ultra-long text lines are dynamically split into multiple sub-text strings, and then the character recognition position is performed. Processing of overlapping characters in adjacent substrings of information.

II. To solve the problem of character recognition in ancient books with poor image quality, close adjacent character structures, and large overlap between upper and lower lines, a Tibetan ancient book string recognition model based on a mixture of residual network and bidirectional long short-term memory network was constructed. The structure of this paper is: the second part introduces the proposed Tibetan ancient book text recognition model, including the introduction of the overall model framework, data set construction and preprocessing, and the string recognition algorithm based on deep learning; the third part is mainly the experimental part, including the description of the data set, evaluation indicators and analysis of experimental results; the fourth part briefly summarizes the results, advantages and innovations of the model in this article.

2. Deep Learning Models for Character Recognition in Tibetan Ancient Books

2.1. Tibetan Ancient Book Recognition Model Framework

In view of the characteristics of Tibetan ancient books and documents, the recognition model framework proposed in this article is divided into a data set construction and preprocessing module, a sliding window-based line recognition module, a deep learning-based string recognition module, and a post-processing of ancient book recognition based on Tibetan spelling check. module, as shown in Fig. 1.

2.2. Dataset Construction and Its Preprocessing

Tibetan language, with a history extending over 1,300 years, has undergone significant evolutionary developments, culminating in a comprehensive and systematic set of spelling rules. The typographic styles within Tibetan scripts are predominantly classified into two categories: 'black gold' and 'black plum'. The 'black gold' style is the primary choice for the fonts used in ancient Tibetan book engravings and printed classics, embodying

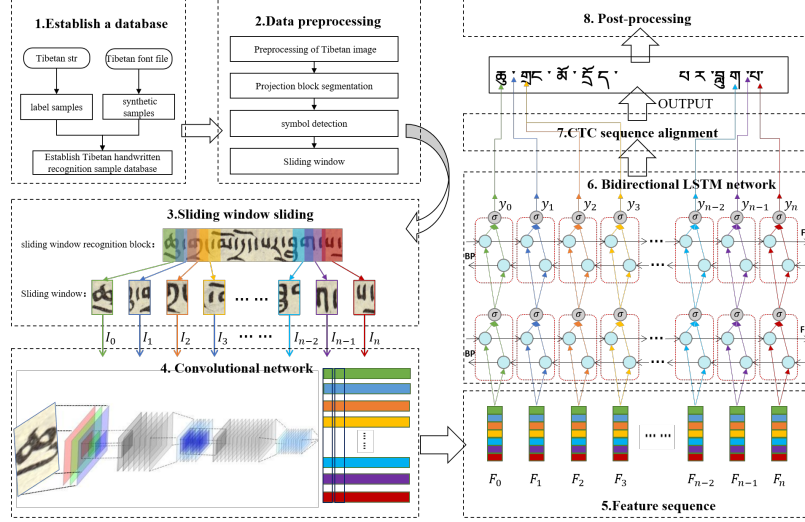


Figure 1: Tibetan ancient book character recognition model framework

the traditional aesthetic and formal presentation of Tibetan script. On the other hand, the 'black plum' font is primarily reserved for handwritten manuscripts and is generally not utilized in the engraving and printing of classical texts.

The foundational structure of the Tibetan script comprises 30 consonant letters and 4 vowel letters, which are arranged in combinations both laterally (left to right) and vertically (top to bottom). The script formations arising from lateral combinations are referred to as 'horizontal', whereas those resulting from vertical combinations are termed 'vertical combination characters'. This intricate system facilitates a rich diversity of script forms and expressions within the Tibetan language.

Ancient Tibetan documents also feature a significant number of transliterations from Sanskrit, showcasing characters that are vertically assembled from Tibetan consonants and vowels. These transliterated characters do not adhere to the standard compositional rules of Tibetan script. Instead, they represent a unique aspect of Tibetan's scriptural heritage, incorporating elements from Sanskrit to enrich its linguistic and cultural tapestry. This fusion of Tibetan and Sanskrit elements within ancient manuscripts highlights the dynamic interchange and integration of cultural and linguistic influences across different traditions, further emphasizing the complexity and depth of Tibetan scriptural studies. The exploration of text recognition in ancient Tibetan manuscripts, particularly those handwritten, confronts several unique challenges detailed in the discussion of Fig. 2. This figure exemplifies the intricacies of abbreviations and Sanskrit-Tibetan transliterations found in these manuscripts, alongside the structural analysis of Tibetan Vertical Concatenation Characters (VCC)[27]. VCCs represent a fundamental component of Tibetan script, consisting of vertical combinations of vowel and consonant letters, distinguished by the use of a specific separator to demarcate syllables, each comprising one or more characters.

Ancient Tibetan texts, which contain a large number of composite characters, such as vertical combinations of vowels and consonants, may extend vertically during the

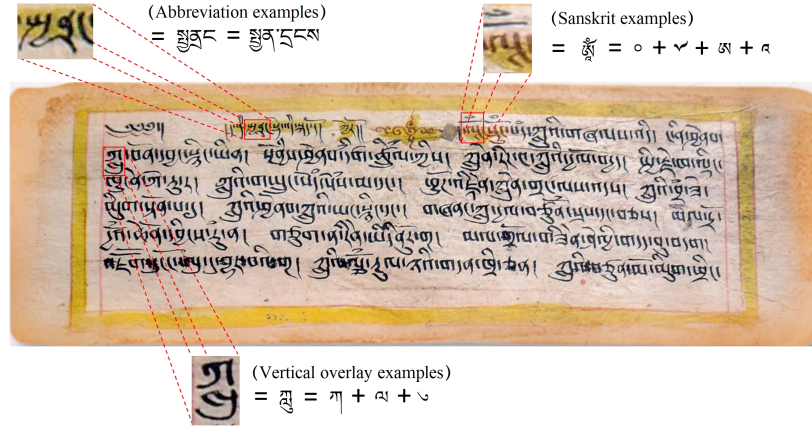


Figure 2: Characteristics of fonts in Tibetan ancient handwritten books

typesetting process, and due to the dense layout of ancient texts, certain characters, especially high or low modifiers, may come close to, or even touch, neighboring lines, resulting in blurring of the text boundaries, and there are also problems such as stains and weathering due to the age of the ancient texts. Problems. These inherent characteristics make the images of Tibetan ancient books show certain features, These include poor image quality with blurred text and significant background noise, minimal spacing between characters, These include poor image quality with blurred text and significant background noise, minimal spacing between characters, resulting in frequent sticking of characters (notably syllable separators and single vertical characters), small line spacing leading to line overlap, and the presence of both Sanskrit and Chinese characters. The uniform width of VCCs within the same text permits their use as a primary unit for recognition.

The character set for recognition encompasses a diverse array of 2013 characters, including 38 non-Tibetan characters (such as Arabic numerals and basic punctuation), 873 Tibetan characters, and 1102 commonly used Sanskrit-Tibetan transliteration characters. This comprehensive character set forms the basis for the subsequent recognition processes discussed in the paper.

Data preprocessing for the recognition of these ancient texts involves several critical steps: scaling images to uniform proportions, normalizing dimensions to 350*48, converting images to grayscale and enhancing them, and utilizing annotation tools to mark the positions of text within images. This annotated data, comprising image files, location annotation files, and text files, ensures that the text extracted from images corresponds accurately with the actual content. Additionally, the generation of synthetic samples through image synthesis technology facilitates the creation of multi-font, single-line samples. This approach not only augments the dataset with a wider variety of text representations but also enhances the robustness and accuracy of the recognition model by training it across a broader spectrum of text characteristics. This meticulous process of data preparation and enhancement is crucial for overcoming the challenges posed by the preservation state and the complexity of ancient Tibetan manuscripts, aiming to achieve precise and reliable text recognition.

2.3. Row Recognition based on Sliding Window

The text recognition process for ancient Tibetan documents presents unique challenges due to the long lines of text, sometimes extending up to 250 characters, and the limitations of the string recognition core which supports a maximum character length of 22. This constraint is based on the character length distribution observed in training samples. Consequently, to accommodate these lengthy lines, the images must be split into multiple strings (blocks) before being processed by the recognition core. This split, however, is complicated by the common issue of characters sticking together, potentially leading to a single character being erroneously divided between two blocks, thereby hampering recognition accuracy.

To mitigate this issue, the paper proposes specific splitting principles aimed at maintaining the integrity of characters during split: I. The aspect ratio of each string (block) must align with the recognition requirements, ensuring that the number of characters within a block does not exceed the capacity of the recognition core, which could degrade accuracy; II. The split process must avoid splitting a single character across two blocks, preserving the character's wholeness for accurate recognition.

To adhere to these principles, the projection method is employed to identify divisible positions that do not result in the splitting of characters. Given the prevalence of noise in images of Tibetan ancient books, such as stains or creases, which can interfere with the projection, a preparatory step involves noise removal. This is achieved by binarizing and denoising the images, followed by calculating all connected domains. The heights of these domains are measured, with the median height serving as a reference (Refheight). If this median is below 20, a default Refheight of 20 is applied based on empirical data. This reference height is then used to discern and eliminate noise by evaluating the size consistency of the connected domains.

In order to realize the above splitting principle, the projection method is used to find all the divisible positions that will not split a character into two parts. We identify inter-character gaps by computing the column-wise sum of pixel values in the projection profile. This approach ensures that split points are accurately positioned within the spaces between characters, thereby preventing errors caused by splitting within characters or at adhesion part between adjacent characters. There is a lot of noise in the images of Tibetan ancient books. In addition to some point noise that can be solved by image processing, there are also some long noises like stains or creases. This type of noise has a greater interference with projection. Therefore, before projection, to remove these noises. Binarize and denoise the Tibetan ancient book images and calculate all connected domains; calculate the heights of all connected domain blocks, and take the median of the heights as the reference height (Refheight). If the median is less than 20, the reference height is set to 20 (20 is the empirical value); based on Refheight, create a rule to determine whether the size of the connected domain is consistent. If not, it is treated as noise and removed; the remaining connected domain blocks are, using the horizontal projection method for split, thus dividing a line into multiple small blocks, and each character only belongs to a certain block. During the split process, we dynamically adjust split thresholds based on the aspect ratio and morphology of characters. This allows the method to accurately identify and avoid mis-splitting, even in cases where minor connections exist between characters. While perfect split cannot be guaranteed, our method significantly reduces the likelihood of errors through strict rules and precise projection analysis. It has

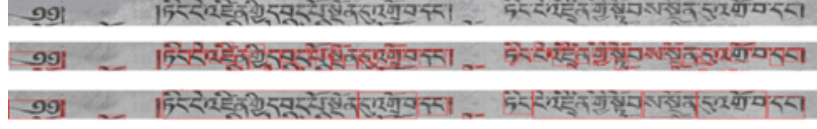


Figure 3: Single line Tibetan projection split

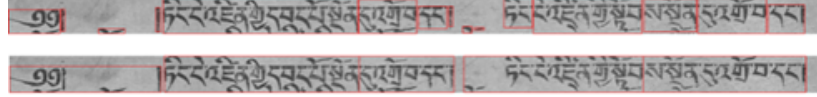


Figure 4: Sliding window to get recognition block

achieved satisfactory results across a wide range of test scenarios. Here we call the divided small blocks projection blocks. Fig. 3 shows the projection split result of a sample, and the result of using the sliding window to extract the recognition block is shown in Fig. 4.

As shown in Fig. 3, the first row is the original grayscale image, and the second and third rows are the connected domain blocks and projection split blocks marked with quadrilaterals on the original image.

As shown in Fig. 4, the first line is the projection block, and the second line is the recognition block. The specific process of sliding window block selection can be divided into three steps: the first step is the calculation of the sliding window parameters; the second step is the sliding of the sliding window; the third step is the adjustment of the identified block boundaries. The two parameter indicators of the sliding window are the aspect ratio (k_window) and the overlap width ($OverLap_Width$), which can be calculated by eq. 1 and eq. 2.

$$k_window = \frac{height_input_layer}{width_input_layer} \quad (1)$$

$$OverLap_Width = k * Refheight \quad (2)$$

In eq. 1, $height_input_layer$ and $width_input_layer$ are respectively the height and width of the image corresponding to the string recognition core input layer. The k value in eq. 2 is 3 based on experience, and $Refheight$ is the reference height.

The sliding window-based line recognition introduced in this section divides the text lines of Tibetan ancient books into multiple small blocks according to the maximum number of characters supported by the string recognition core, providing a basis for subsequent string recognition.

2.4. Tibetan Ancient Book Document String Identification Model Framework

String recognition technology is the core algorithm for text recognition in Tibetan ancient books. To solve the problem of identifying ultra-long text lines in Tibetan ancient books, this paper designed a string recognition method based on deep learning. The string recognition framework is shown in Fig. 5.

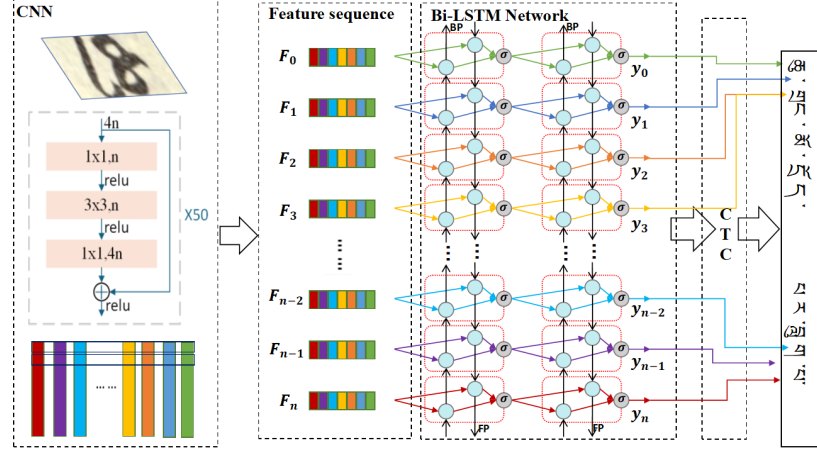


Figure 5: String recognition framework based on deep learning

As shown in Fig. 5, in string recognition based on deep learning, the image is resized and normalized into a 3-channel tensor of size 224×224 . The tensor first passes through a single convolutional layer followed by a pooling layer, and is then processed by the ResNet model. Finally, the output feature maps are flattened into a one-dimensional vector to obtain the feature extraction results, where n represents the dimensionality of the features and varies depending on the number of layers.

The calculation is defined in eq. 3 :

$$y = F(x, \{w_i\}) + x, \quad (3)$$

where x denotes the input features to the residual block, $F(x, \{w_i\})$ represents a series of transformations (e.g., convolution, Batch Normalization [33], and ReLU activation) with weights w_i , and y represents the output of the residual block. Bi-LSTM is used to learn to associate sequences and predict label distribution; the transcription layer uses CTC for sequence alignment and outputs prediction results. Due to the serious adhesion of characters in ancient Tibetan books, string recognition uses Bi-LSTM, which is a recursive neural network that can learn long-term dependency information and is suitable for processing sequence data problems [28]. Using LSTM to solve the string recognition problem of text images is to treat text images as sequential events [29]. The Bi-LSTM layer uses the extracted features as sequential input, enabling it to capture contextual information from both past and future directions. At each time step, the forward and backward LSTM outputs are defined in eq. 4 and eq.5:

$$h(t) = f(Ux(t) + Wh(t-1)), \quad (4)$$

$$y(t) = g(Vh(t)), \quad (5)$$

U , W , and V denote the input, recurrent, and output weights, respectively. These weights are optimized during training to adjust how the input and previous state contribute to the current state. $f(z)$ and $g(z)$ represent the sigmoid and softmax activation functions, respectively, and are defined in eq. 6 and eq. 7 :



Figure 6: CTC Merge Sequence Mechanism

$$f(z) = \frac{1}{1 + e^{-z}}, \quad (6)$$

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}. \quad (7)$$

According to the maximum string length supported by the recognition core, the original 350*48 compression is performed as 44*6 (the height is reduced by 8 times and the maximum string length supported by the recognition core is 2 times), that is the input Timestep of the LSTM network is 44, and the data dimension output by Timestep is 2013, which indicates that the timestep corresponds to the probability of 2013 categories. Each timestep takes an optimal category, and by merging these 44 categories, the string recognition results corresponding to the input image can be obtained. The data set used for string recognition training is divided into synthetic samples and actual labeled samples. First, synthetic samples are used to pre-train a Tibetan string recognition model to extract better shallow features. As the actual labeled samples increase, a finetune training sample set can be established. This sample set is a combination of synthetic samples and real samples in a certain proportion. Among them, finetune refers to continuing to train the current sample based on a pre-trained model, so that there is no need to retrain the model. Through this iterative approach, the performance of the recognition core is continuously improved, and a model with ideal recognition effects on both synthetic samples and real samples is finally obtained.

Finally, CTC simulates the dynamic programming process by introducing blank classes and mapping rules [30]. The CTC module uses Softmax and a fully connected network to convert the prediction results of each sequence into a label sequence and uses the CTC merging sequence mechanism to predict the results of the sequence output by LSTM, ultimately realizing the fusion training of CNN and RNN. The CTC alignment process is shown in Fig. 6.

As shown in Fig. 6, combine consecutive repeated characters when outputting, like that from t7 to t10, the correspond to Tibetan character is “ba”, means that “ba ba ba ba” will merge into “baba” which means four repeated Tibetan characters are merged

into one character. This merging mechanism is obviously problematic. In order to avoid merging duplicate characters that already exist, CTC [31] introduced a whitespace mechanism, using “-” Character representation is inserted between repeated characters, which solves the problem of merging errors caused by repeated characters.

In all, this paper uses the model CRNN, which is a fusion of CNN and RNN, as a model for Tibetan string recognition. This model is good at identifying text sequences where character adhesion makes split difficult.

3. Experimental Results

3.1. Datasets

Currently, there is no public image annotation data or test data for Tibetan ancient books and documents. To verify the effectiveness of the model proposed in this study, this paper constructed an image annotation data set for Tibetan ancient books and documents from Tibet University.(Tibetan Ancient Books Dataset, TABD-2021[32]),This dataset is divided into training data and test data. The training data set includes: 3,000 images of modern Tibetan printed documents, 3,000 images of ancient Tibetan woodblock documents, 3,000 images of handwritten images of ancient Tibetan books, and 4,000,000 synthetic images. The synthetic images are all small single line images. The test data set includes: 100 images each of modern prints, woodcuts, and handwritten documents, for a total of 300 images. The source of the original image is obtained through borrowing documents from the Tibetan Library, the Academy of Social Sciences, the University of Tibetan Medicine, some monasteries, and other institutions, and is scanned by a scanner. The documents were scanned under uniform lighting conditions to minimize the impact of external variables, thereby demonstrating the performance of the model more effectively under standardized conditions.The text corresponding to the image is obtained by manual annotation, as shown in Tab. 1.

Table 1: Dataset Details

Category	Modern Typography	Ancient Woodblock	Handwritten Ancient Books
Real-Dataset	3000	3000	30000
Synthetic	0	4000000	0
Test	100	100	100

As shown in Tab. 1,The units of all quantities are in sheets, the real samples are all scanned Tibetan document images, and the synthetic images are all single-line Tibetan images, which are images synthesized based on the layout, background, font and other characteristics of the Tibetan document. we used both real and synthetic data for training in the “ancient woodblock” category, while the “handwritten ancient books” and “modern typography” categories were trained exclusively on real data. This design choice was made because the “ancient woodblock” category involves challenges such as character ligatures and Sanskrit-Tibetan transliterations, making it more difficult to recognize. However, the amount of real data available for this category is insufficient.To address this, we augmented the “ancient woodblock” dataset with synthetic data, allowing our model to achieve better generalization performance during training. Both real and synthetic

data were used for training in the “ancient woodblock” category, whereas only real data was used in the “handwritten ancient books” and “modern typography” categories. The test data are Tibetan images that have not been added to the training. 100 images are selected from each of modern printed documents, Tibetan ancient woodblock documents, and Tibetan ancient handwritten documents. The test samples cover all characters in the character set.

3.2. Evaluation Indicators

In order to verify the validity of the experiment, this article adopts the macro-averaging calculation method (Macro-averaging), as shown in eq. 8 - eq. 10.

$$Macro_P = \frac{1}{n} \sum_{i=1}^n P_i, \quad (8)$$

$$Macro_R = \frac{1}{n} \sum_{i=1}^n R_i, \quad (9)$$

$$Macro_F = \frac{1}{n} \sum_{i=1}^n F_i, \quad (10)$$

As appeared in eq. 8, $Macro_P$ is the accuracy, and its formula is $P = \frac{TN}{LN+MN}$, TN represents the number of correctly recognized characters, LN represents the total number of characters labeled in the test set, and MN represents the number of multiple recognized characters; $Macro_R$ in eq. 9 is the recall rate, and the calculation formula can be expressed as $R = \frac{TN}{LN}$; As appeared in eq. 10, $Macro_F$ is F , and the calculation formula can be expressed as $F = \frac{2 \times P \times R}{P+R}$.

3.3. Comparative Experiments & Analysis

Currently, for Tibetan recognition, the publicly available software mainly includes TH-OCR, jointly developed by the Department of Electronic Engineering of Tsinghua University and Northwest University for Nationalities, and UTibet-OCR-2008, developed by the Tibetan Information Technology Engineering Research Center of the Ministry of Education of Tibet University. However, these systems can only recognize modern printed fonts and cannot be applied to the recognition task of ancient Tibetan books. Therefore, this article selects 300 images from the TABD-2021 data set as test data to identify modern printed fonts, ancient woodcuts, and ancient handwritten texts. Do a test on. Fig. 7 shows the test results of modern print recognition accuracy.

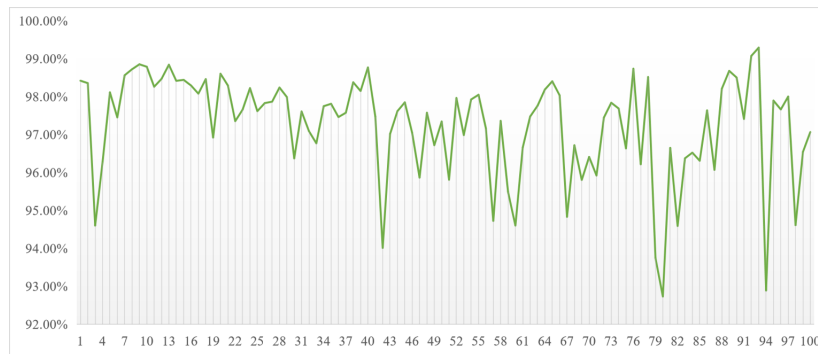


Figure 7: Tibetan Modern Print Recognition Accuracy

As shown in Fig. 7, the recognition accuracy results of 100 modern printed test samples are given. It can be seen that on some samples, the highest accuracy rate reaches 99.30%, and the lowest accuracy rate also reaches 92.74%.

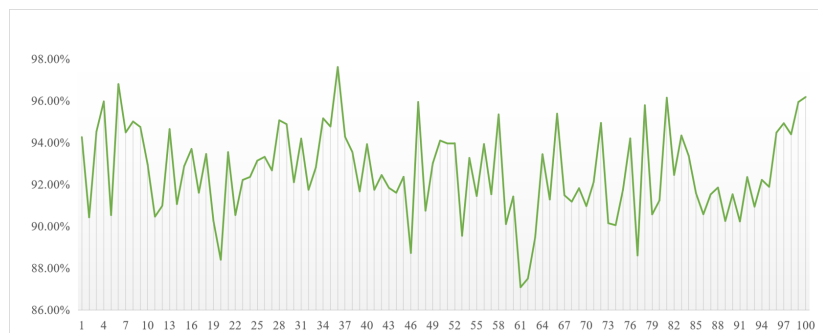


Figure 8: Recognition accuracy of woodblock versions of Tibetan ancient books

As shown in Fig. 8, the test sample of 100 ancient book woodcuts was given, and the highest and lowest accuracy rates were 97.62% and 87.08% respectively.

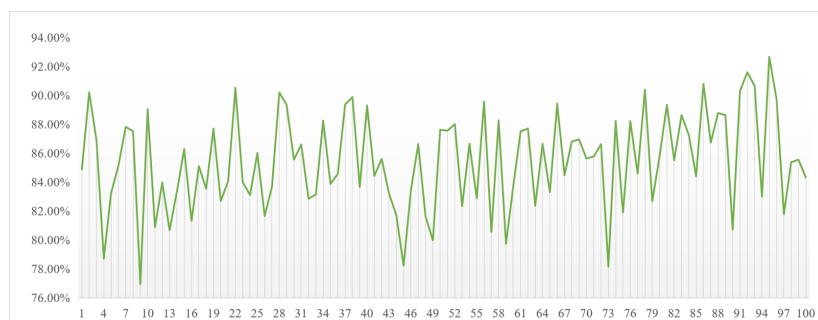


Figure 9: Accuracy of handwriting recognition in Tibetan ancient books.

As shown in Fig. 9, the recognition accuracy results of 100 ancient handwritten book test samples are given. It can be seen that the highest accuracy rate reached 92.67%, and the lowest accuracy rate reached 76.96%. According to the evaluation indicators, the values of $Macro_P$, $Macro_R$ and $Macro_F$ can be obtained, as shown in Tab. 2.

Table 2: Experimental results

Test sample	Macro_P	Macro_R	Macro_F
modern typography	97.27%	98.42%	97.84%
ancient woodblock	92.60%	94.06%	93.32%
handwritten ancient books	85.47%	86.43%	85.95%

As shown in Tab. 2, it can be found that the indicators of this model have reached the optimal level in the Tibetan modern print recognition task, with the macro average accuracy, recall rate, and F value reaching 92.27%, 98.42%, and 97.84% respectively; in Tibetan The ancient book woodblock recognition task reached 92.60%, 94.06%, and 93.32% respectively; the Tibetan ancient handwritten book recognition task reached 85.47%, 86.43%, and 85.95% respectively. The recognition performance of the handwritten book is compared with the first two. lower. Through analysis, it was found that in handwritten books, there are multiple characters written across characters and abbreviations, which leads to a decline in recognition performance. There are many Sanskrit and Tibetan transliteration characters in the woodcut editions of ancient Tibetan books. Due to the age, the Tibetan separators are unclear or disappear. These problems are still not effectively solved by this model.

Since the TH-OCR and UTibet-OCR-2008 recognition systems have poor recognition performance for Tibetan ancient books and documents, we only compare the recognition performance of modern prints on the test set. The comparison results are shown in Tab. 3.

Table 3: Comparative Experiments

Models	Macro_P	Macro_R	Macro_F
TH-OCR	85.26%	86.54%	85.90%
UTibet-OCR-2008	86.32%	87.06%	86.69%
Our models	97.27%	98.42%	97.84%

As shown in Tab. 3, through comparative experiments, it is found that in modern print recognition, the average macro accuracy of TH-OCR and UTibet-OCR-2008 systems reached 85.26% and 86.32% respectively, and the average macro accuracy of the text model reached 97.27%, compared with the first two systems, its macro average accuracy increased by 12.01 and 10.95 percentage points respectively.

For comparative analysis, we used Yuncong OCR [34], Lingyun OCR [35], and Tesseract OCR [36], which are open-source OCR tools developed by Lingyun Technology, Yun-smart Technology, and Google, respectively. The test samples were submitted to the

public interfaces of each model, with the language set to Tibetan, and the recognition which are open-source OCR tools developed by Lingyun Technology, Yunsmart Technology, and Google, respectively. results were retrieved. Using the ocreval[39] testing software, we calculated the average macro accuracy of the three test categories based on their corresponding annotations. The results are summarized in Tab. 4 :

Table 4: Macro-P Performance Comparison Across Different Datasets

Test sample	modern typography	ancient woodblock	handwritten ancient books
YuncongOCR[34]	96.83%	65.53%	68.75%
LingyunOCR[35]	96.29%	70.14%	3.87%
TesseractOCR[36]	86.40%	48.85%	15.18%
Our models	97.27%	92.60%	85.47%

As shown in the Tab. 4, the performance of the models varies significantly across test texts, with our model outperforming all others on every test dataset. All four models demonstrate strong Macro Accuracy on modern typography. In particular, our model achieves a precision of 97.27% on it. outperforms the other models, showing an advantage in dealing with standardised texts. Ancient woodcut texts suffer from glyph variability and poor print quality, and the performance of the models varies greatly in targeting their recognition. Our model still performs well on ancient woodblock texts, with a Macro Accuracy of 92.60%, obtaining an improvement of 22.46% to 43.75% compared to the other three models. It shows its strong ability in dealing with the noise of ancient texts. In the recognition of handwritten ancient books, our model achieves a Macro Accuracy of 85.47% on it, obtaining an improvement of 16.72%-81.6%, which is a good performance in this challenging task, proving that our method is able to cope with the diversity and irregularity of handwritten books better, especially when dealing with low-quality and blurred antique manuscripts, and still maintains a high accuracy.

3.4. Qualitative Analysis

Visualization To further demonstrate the superior performance of our model, we provide visual results for handwritten ancient books and woodblock prints.

	handwritten ancient books			ancient woodblock		
Yuncong OCR	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་
Lingyun OCR	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་
Tessact OCR	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་
Ours	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ཧུང་ཅུང་འཕྲུལ་འཕྲུལ་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་	ལུང་ཡུང་ལུང་ལུང་ལུང་ལུང་
	(1)	(2)	(3)	(4)	(5)	(6)

Figure 10: Visual comparison of recognition results

As shown in Fig. 10, our recognition results achieve superior performance compared to other baseline models. For handwritten examples 1 and 2 of ancient texts with connected characters, our model successfully splits and recognizes the connected characters, effectively reducing recognition errors caused by character connections. In handwritten example 3, which contains abbreviated handwritten characters, our model achieves accurate recognition results, whereas other models either fail to recognize these characters or produce incorrect results.

In ancient woodblock examples 4 and 5, challenges such as ink diffusion and severe stains due to the age of the texts are present. Under these issues, our method maintains high recognition accuracy, demonstrating significant robustness in handling noise interference and strong adaptability to complex scenarios. Moreover, our method shows high adaptability to different fonts and layouts.

Failure cases and discussion Recognition failures of our model on Tibetan ancient woodblock and handwritten ancient books are also presented.

	Image Screenshot	Output	Ground Truth
(1)		གཏིང་ལ་དག་ཐོང་།	གཏིང་ལ་དག་ཐོང་།
(2)		དུས་འདྲི་དྲ་མཐའ་གྲུ་གྲུ་།	དུས་འདྲི་དྲ་མཐའ་གྲུ་གྲུ་།
(3)		ཁོ་ལོ།	གཞོན་པ།

Figure 11: Examples of woodcut and handwritten failures

As shown in the Fig. 11, Example 1 is an example of woodcut recognition error, which is caused by excessive noise due to stains, and the model is difficult to identify the character features. It shows that the robustness of the model needs to be further improved when dealing with the complex background caused by pollution or aging in woodcuts, and also reflects the necessity of data preprocessing and noise reduction algorithm optimisation.

The high variability in the shapes of handwritten characters, particularly the significant deformation or abbreviation of certain characters, adversely affects recognition accuracy. For example, in instance 2, cursive handwriting-induced deformation caused the model to misclassify the character 'ཞ' as 'ཏ' and 'ཏ' as 'ཏ'. This highlights the challenges the model faces in differentiating specific handwritten characters (e.g. 'ཞ' and 'ཏ') due to individual writing habits. It underscores the necessity of enhancing the feature extraction mechanism to improve sensitivity to both local character details and overall structural variations in handwritten characters. In instance 3, when processing the handwritten character 'ཁོ', its upper handwritten extension disrupted the split strategy. Specifically, as the method split based on vertical text layout units, the extended part of the character was erroneously split into three separate character regions above it, leading to recognition errors. This demonstrates that, while the method is robust for standard-layout ancient texts, further refinement is needed to better handle the diversity of handwritten character shapes.

Conclusion

This paper introduces a novel approach to address the challenges associated with text recognition in ancient Tibetan manuscripts, characterized by poor image quality, extended text lines, and characters that are closely spaced or overlapping. The proposed solution encompasses a sliding window-based line recognition technology coupled with a deep learning-based string recognition methodology. Experimental validations of this technology have yielded promising results: a macro-average accuracy of 97.27% in recognizing printed text from modern Tibetan publications, 92.60% in character recognition within woodblock editions of ancient Tibetan texts, and an impressive 85.47% in handwritten editions, comparisons were made with currently popular Tibetan text recognition models on three different types of datasets, and our models all achieved the best recognition results, the visualization results demonstrate that our model effectively handles character adhesion and severe image noise. Marking a significant advancement in the domain of ancient Tibetan text recognition research both domestically and internationally. The developed system has been launched for a public free trial, contributing to the digital preservation and exploitation of Tibetan ancient manuscripts. Future endeavors will focus on addressing the specific challenges outlined in Section 2.3 of this paper, including the recognition of Tibetan social media texts and the identification of multi-font texts within complex scenes, leveraging the foundational technology established through this research.

Author Contributions

Rinchen Dongrub focused on model training and algorithm optimization, the introduction of the latest deep learning technologies has significantly improved the accuracy and efficiency of text recognition. Dorje Tashi gathered a vast collection of Tibetan ancient books' images from diverse sources to build a high-quality dataset. Thupten Tsering: Proposed a line recognition method based on sliding windows to tackle challenges like hidden texts and excessively long lines in ancient books. Nyima Tashi Offered professional guidance in research design, data analysis, and significant direction during the writing phases, shaping the manuscript's scholarly narrative. YU Yong-bin provided expert guidance on the formulation of the conceptualization and the structure of the research framework. CHENG-Jian participated in the implementation of the approach, and provided technical support.

Acknowledgements

We would like to extend our sincere gratitude to the authors from The Tibet Library, The Academy of Social Sciences, and The Tibetan Medical University, for their participation in the creation of Tibetan Ancient Books Dataset. This work is supported in part by the National Science and Technology Major Project (2022ZD0116100), and in part by the Province and Ministry Jointly Constructed Tibetan Language Intelligent Information Processing and Application National Key Laboratory Open Topic (2023-Z-006), and in part by the National Natural Science Foundation of China Project (62306158).

References

- [1] Fu, M.: Problems of the Creation and Reform of Scripts by China's Ethnic Minorities. *Minzu Yanjiu* (Ethnic Studies), 9–17 (1979).
- [2] Xianba.: Research on the Editions of Ancient Tibetan Texts: Focusing on the Printed Copies of Ancient Tibetan Books. *Tibet Studies*, 99–110 (2016).
- [3] Rowinski, Z., Keutzer, K.: Namsel: An Optical Character Recognition System for Tibetan Text. *Himalayan Linguistics*, 15(1), 1–20 (2016).
- [4] Kojima, M., Nunomiya, C., Kawamura, T.: Recognition of Similar Characters by Using Object Oriented Designed Printed Tibetan Dictionary. *Ipsj Journal*, 36, 2611–2621 (1995).
- [5] Kojima, M., Kawazoe, Y., Kimura, M.: Automatic Recognition of Wooden Blocked Tibetan Image Character by Using Object Oriented Design. *Ipsj Sig Notes*, 98, 39–44 (1998).
- [6] Wang, W., Ding, X., Chen, L., et al.: Research on the Recognition of Printed Modern Tibetan Script. *Computer Engineering*, 29(3), 37–38, 94 (2003).
- [7] Wang, H., Zhao, N., Deng, G.: Preprocessing for Tibetan Script Recognition. *Computer Engineering*, 09, 93–96 (2001).
- [8] Pu, C.: Research on Tibetan Script Recognition Technology. Tibet University, (2008).
- [9] Wu, G., De, X., Huang, H.: Technology for Recognizing Printed Tibetan Script. *Journal of Qinghai Normal University (Natural Science Edition)*, 01, 32–37 (2006).
- [10] Wang, H., Ding, X.: A Method for Recognizing Multi-Font Printed Tibetan Characters. *Computer Engineering*, (13), 18–20 (2004).
- [11] Qi, K.: Research on Rule-Based Post-Processing for Tibetan Script Recognition. *Journal of Northwest Minzu University (Natural Science Edition)*, (04), 33–36 (2003).
- [12] Zhi, X., Gao, D.: Research on the Detection Method for Handwritten Multi-Font Tibetan Ancient Texts. *Plateau Science Research*, 6(02), 89–101 (2022).
- [13] Rinchen, D.: Research on the Recognition of Tibetan Ancient Texts from Woodblock Prints Based on Deep Learning. Tibet University, (2021).
- [14] Nyima, T.: Research and Application of Tibetan Script Spelling Formal Languages and Their Automata. Science Press, (2016).
- [15] Nyima, T.: Research on Tibetan Spelling Formal Language and Automata with Application. Springer, Singapore, (2018).
- [16] Wang, D.: Research on Feature Extraction for Online Handwritten Tibetan Recognition Based on Statistical Structures. Northwest Minzu University, (2012).
- [17] Cuo, Y., Tashi, N., Liu, Z., et al.: Layout Analysis of Tibetan Historical Documents Based on Deep Learning. *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*, 96–101 (2019).
- [18] Rinchen, D., Thupten, T., Nyima, T.: Research on the Synthesis Method of Training Data for Tibetan Ancient Text Recognition. *Plateau Science Research*, 5(03), 84–91 (2021).
- [19] Lu, Y.: Research on Super-Resolution Reconstruction of Tibetan Ancient Document Images. Northwest Minzu University, (2023).
- [20] Liu, Y.: Research on Multi-Line Text Recognition Technology for Tibetan Ancient Texts. Tibet University, (2023).
- [21] Ding, H.: Research on Compression and Acceleration of Sequence Recognition Models Based on Deep Learning. University of Science and Technology of China, (2020).
- [22] Villameriel, S., Costello, B., Dias, P., Giezen, M., Carreiras, M.: Language modality shapes the dynamics of word and sign recognition. *Cognition*, 191, (2019).
- [23] Khalifa, M., Shaalan, K.: Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks. *Computer Speech and Language*, 58, (2019).
- [24] Zhou, M., Cairang, X., Long, C., Yan, X.: Research on Tibetan Handwritten Digits and Letters Recognition Based on Convolutional Neural Networks. *Journal of Qinghai Normal University (Natural Science Edition)*, 35(04), 34–39 (2019).
- [25] Cairang, D.: Research on Tibetan Uchen Script Recognition Technology Based on Syllable Segmentation. Qinghai Normal University, (2020).
- [26] Rinchen, D., Nyima, T.: Research on Tibetan Script Recognition in Natural Scenes Based on Deep Learning. *Plateau Science Research*, 3(04), 96–103 (2019).
- [27] Nyima, T.: Research and Application of Tibetan Script Spelling Formal Languages and Their Automata in the Series of Academic Works on Information Science and Technology. Science Press, Beijing, (2016).
- [28] Sulpizio, S., Navarrete, E.: Outgroup faces hamper word recognition. *Psychological research*, (2019).

- [29] Wang, Y.: Research on Tibetan Printed Character Recognition Technology Based on Improved Coarse Grid. Jiangsu University of Science and Technology, (2010).
- [30] Li, Y., Wang, Y., Liu, Z.: Research on the Technology of Tibetan Printed Character Recognition. Journal of Nanjing University (Natural Sciences Edition), 48(01), 55–62 (2012).
- [31] He, X.: Research on the Recognition of Continuous Handwritten Digit Strings Based on Improved BP Neural Network. Beijing University of Technology, (2014).
- [32] Rinchen Dongrub.: TABD-2021. <https://github.com/rindon520/-TABD-2021>, (2024).
- [33] Ioffe, S.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015).
- [34] Yuncong Intelligence Technology Co., Ltd.: Yuncong Smart Ancient Books Digitalization Platform. Retrieved November 2, 2024, from <https://www.yuncong-ai.com/> (2023).
- [35] Jietong Huasheng: Lingyun - Empowering industries, sharing AI future. Retrieved November 2, 2024, from <https://www.aicloud.com/> (2021).
- [36] Smith, R.: An overview of the Tesseract OCR engine. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, vol. 2, pp. 629–633 (2007). IEEE.
- [37] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).
- [38] Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, vol. 18, no. 5-6, pp. 602–610 (2005).
- [39] Santos, E. A.: OCR evaluation tools for the 21st century. In Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers), pages 23–27, Honolulu. Association for Computational Linguistics (2019).