electronics

MDPI

Article

Offline Mongolian Handwriting Recognition Based on Data Augmentation and Improved ECA-Net

Qing-Dao-Er-Ji Ren [1], Lele Wang [1,*], Zerui Ma [1] and Saheya Barintag [2]

[1] School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China; renqingln@imut.edu.cn (Q.-D.-E.-J.R.); 20221800122@imut.edu.cn (Z.M.)
[2] School of Mathematics Science College, Inner Mongolia Normal University, Hohhot 010028, China; saheya@imnu.edu.cn
* Correspondence: wangll@imut.edu.cn

Abstract: Writing is an important carrier of cultural inheritance, and the digitization of handwritten texts is an effective means to protect national culture. Compared to Chinese and English handwriting recognition, the research on Mongolian handwriting recognition started relatively late and achieved few results due to the characteristics of the script itself and the lack of corpus. First, according to the characteristics of Mongolian handwritten characters, the random erasing data augmentation algorithm was modified, and a dual data augmentation (DDA) algorithm was proposed by combining the improved algorithm with horizontal wave transformation (HWT) to augment the dataset for training the Mongolian handwriting recognition. Second, the classical CRNN handwriting recognition model was improved. The structure of the encoder and decoder was adjusted according to the characteristics of the Mongolian script, and the attention mechanism was introduced in the feature extraction and decoding stages of the model. An improved handwriting recognition model, named the EGA model, suitable for the features of Mongolian handwriting was suggested. Finally, the effectiveness of the EGA model was verified by a large number of data tests. Experimental results demonstrated that the proposed EGA model improves the recognition accuracy of Mongolian handwriting, and the structural modification of the encoder and coder effectively balances the recognition accuracy and complexity of the model.

Keywords: attention mechanism; character recognition; data augmentation; neural network

Citation: Ren, Q.-D.-E.-J.; Wang, L.; Ma, Z.; Barintag, S. Offline Mongolian Handwriting Recognition Based on Data Augmentation and Improved ECA-Net. *Electronics* **2024**, *13*, 835. https://doi.org/10.3390/electronics13050835

Academic Editor: Daniel Riccio

Received: 18 January 2024
Revised: 10 February 2024
Accepted: 20 February 2024
Published: 21 February 2024

1. Introduction

Writing, which records human history and inherits human civilization, is a unique skill and cultural symbol of human beings. In the Inner Mongolia region of China, the traditional Mongolian script used in Inner Mongolia is the Uighur-Mongolian script and plays an important role in the local historical record and cultural inheritance. Digitizing Mongolian handwriting in bulk using character recognition technology is an important way for Mongolian culture to keep pace with the times.

Training character recognition models by using deep neural networks is a common practice in the current character recognition field, and it has a significant effect on Chinese and English recognition. The LeNet5 CNN model proposed by Lecun et al. [1] in 1998 is regarded as an early classical model in the field of character recognition. After data augmentation, the LeNet5 CNN recognized handwriting in the MNIST dataset with 99.2% accuracy. In 2011, scholars from IDSIA Labs used GPUs to train CNNs, opening the application of neural networks in large-class (1000-class) Chinese handwriting recognition. Cire et al. [2] integrated multiple CNNs with various input scales and trained the model with the NIST SD19 dataset (having 800,000 samples), achieving a recognition accuracy of 89.12%. The recognition accuracy on the MNIST dataset was even better, reaching 99.72%. In 2015, Shi et al. [3] used CNN to extract the entire input image features and then

converted the extracted feature map into a feature sequence input to RNN to predict the output sequence. They calculated the sequence output probability with the connectionist temporal classification (CTC). This end-to-end trainable neural network based on image sequence recognition is a classical convolutional recurrent neural network (CRNN). CRNN not only has outstanding achievements in popular character recognition such as Chinese and English script, but also shows excellent recognition performance in various ethnic minority character recognition tasks. In 2021, CRNN was applied to scanned Uyghur character recognition [4], with a recognition accuracy of more than 90%. In the same year, CRNN was also used in recognizing the script of The Book of Changes [5] and Tibetan-Chinese bilingual text [6], and both achieved reasonable recognition results. In 2020, the Baidu Flypaddle team proposed a lightweight character recognition system: PP-OCR [7]. The overall size of the system model is only 3.5 MB and it can recognize 6622 Chinese characters and 63 alphanumeric symbols. After training, the model also showed good performance in character recognition for several other languages (French, Korean, Japanese, and German). Since 2020, researchers have extended the "transformer" to computer vision tasks. In 2020, Dosovitskiy et al. [8] proposed a Vision Transformer (ViT), which applies a standard Transformer directly to images, with the fewest possible modifications, inspired by the Transformer scaling successes in NLP. In addition, ViT has reached or exceeded the latest level on many image classification datasets, showing superior performance. In 2021, Liu et al. [9] of Microsoft Research Asia proposed a new type of vision transformer, Swin Transformer, that can be used as a general-purpose backbone network for computer vision. Experimental results have shown that the self-attention mechanism based on the shifted window introduced in the algorithm is an effective method to solve the visual problem. In 2022, Riaz Nauman et al. [10] proposed the CNN-transformer architecture to solve the problem of offline Urdu handwriting recognition. The effectiveness of the proposed method was verified on the public NUST-UHWR dataset. In 2023, Dan Yongping et al. [11] introduced the particle swarm optimization method into the design of CNN handwritten Chinese character recognition, which reduced redundant calculations in the network and achieved good experimental results.

Traditional Mongolian is a kind of script that is written from top to bottom, and the handwritten Mongolian script has the characteristics of diverse styles and flexible glyphs, and it is difficult to divide the morphemes. In 2017, Fan et al. conducted research on Mongolian handwriting recognition based on grapheme segmentation and implemented a handwriting recognition system by using a small-scale font library in the HTK (hidden Markov model toolkit) environment [12]. The experimental results showed that short graphemes have better performance than long graphemes. In recent years, Mongolian handwriting recognition based on whole words has begun to emerge. In 2019, Liu [13] proposed a sequence-to-sequence offline handwriting Mongolian whole-word recognition model with an attention mechanism. On large-scale datasets, the recognition accuracy of the model reached 81.56%. In 2020, Wei [14] proposed an end-to-end handwritten Mongolian whole-word recognition method and the experimental results showed that the method is superior to the traditional ones based on morpheme segmentation. This method not only achieved the optimal recognition effect at that time, but also alleviated the out-of-vocabulary (OOV) problem. In 2021, Yang [15] built an online Mongolian handwriting recognition cloud service system, providing online Mongolian handwriting recognition services via a service interface.

Data augmentation (DA) is an important means to improve the training effect of neural network models by augmenting the training set. DA algorithms for character recognition research have appeared for Chinese, English, and other languages. With the application of machine learning and neural networks in many fields, DA technology has also been greatly developed. In 1998, Lecun et al. [1] developed data augmentation for the MNIST database by using various operations, such as scaling, rotation, and stretching, to effectively improve the recognition accuracy of the character recognition model. Experimental results have shown that the elastic deformation data augmentation method [16] proposed for sequence

character recognition can help improve the efficiency of character recognition models. In 2017, Zhong et al. [17] proposed a random erasing data augmentation (REDA) algorithm for the target detection task. By erasing random areas in the target image, random erasing data augmentation algorithms can improve the robustness against occlusion and realize data augmentation. In 2021, Han et al. [18] used generative adversarial networks (GANs) to enhance the data of ancient characters, reconstruct the region-specific information of the character, and improve the ability of the model to extract image features. It can be seen that in the field of character recognition, it is theoretically feasible to augment the data of character images for training.

The abovementioned research indicates that the use of deep neural networks to achieve Mongolian handwriting recognition has a sufficient theoretical basis, but directly pouring traditional Mongolian corpus into the character recognition models for popular languages such as English and Chinese may not be feasible, since the recognition performance may be very poor. Therefore, we need to fully explore the unique features of traditional Mongolian on the basis of sufficient handwritten character corpus and build a model suitable for Mongolian handwritten script. Moreover, the traditional Mongolian handwriting recognition research started late, and the public datasets are relatively scarce; therefore, it is imperative to use the DA algorithm to augment the Mongolian handwriting recognition dataset.

In summary, to overcome the obstacles to the current Mongolian handwriting recognition method, such as imperfect data foundation and low prediction accuracy, this paper proposes an improved handwriting recognition model suitable for Mongolian handwriting characteristics. The main contributions of this paper are as follows: horizontal wave transformation (HWTDA) and random erasing data augmentation (REDA) algorithms were studied and used to augment Mongolian handwritten data. In model training, ECA, GRU, and Attn modules were used to encode and predict Mongolian handwriting to fully explore the local and global features of character data to improve the prediction performance of our model.

2. An Improved Dual Augmentation Algorithm

The current mainstream handwriting recognition methods, especially the methods represented by deep neural network models, need to be based on a large amount of data, and small-scale data sets cannot optimize the parameters of the training model. Research shows that when the amount of data in the dataset is small, selecting the appropriate data augmentation method can improve the efficiency of model training to a certain extent [19–22].

Random erasure data augmentation algorithm is a commonly used data augmentation method that can generate more training data by randomly erasing a part of the input image. However, this algorithm may lose some key information, resulting in the performance degradation of the model. Therefore, we modified this algorithm so that it can better adapt to the challenges related to Mongolian handwriting recognition.

The DA method should be selected based on the characteristics of the Mongolian handwritten script. First, the traditional Mongolian script consists of letters, and each letter exhibits different beginning, middle, and ending writing formats in words. Therefore, random clipping data augmentation will destroy the Mongolian format and have a negative impact on model training. Second, traditional Mongolian script follows a fixed writing rule, with letters going from top to bottom and lines running from left to right. Therefore, the samples formed using simple rotation, flipping, etc. are not in line with real ones and are difficult to function. Third, traditional Mongolian script belongs to the phonetic scripts, the characters are seamlessly connected, the position of the characters is relatively free when writing, and the phenomenon of consecutive strokes and omissions is quite common, which provides room for selecting DA algorithms. Finally, in daily life, handwritten characters often have some writing stains due to poor writing habits or low writing instrument quality, and if the model cannot improve the stain inclusiveness, it will inevitably affect its practical performance.

Based on the above practical requirements and the analysis of Mongolian handwriting recognition datasets, the augmentation of Mongolian handwriting data is achieved from the following two aspects:

(1) From the perspective of handwriting image morphology, the HWT algorithm is used for augmenting the Mongolian handwriting data. The algorithm is softer and more flexible in processing the sample of the raw image. The obtained image samples with complete character structure follow the traditional Mongolian writing format, and the augmentation effect is diverse. Moreover, it can further enhance the details of the image and reduce the information lost due to the erasure operation, thereby enhancing the robustness of the model.

The normalized original image is used to perform horizontal wave transformation in the HWTDA algorithm, which can ensure that the parameters are universal for the image. The direction, range, and amplitude of the transformation are determined by parameter values. HWT can change pixels in an image into a horizontal sine wave shape with a given amplitude and frequency; therefore, the conversion parameters, i.e., the conversion frequency N and the magnitude R, need to be specified in advance. Each point in the image follows the transformation rules of Equations (1) and (2).

$$T(x) = x \tag{1}$$

$$T(y) = y + (\pm)R * \sin(N\pi * y) \tag{2}$$

From the above formula, HWT does not produce any distortion on the x-coordinate of a pixel. In Equation (2), the parameter R is used to control the amplitude of the sine wave, N is used to control the frequency of the transformation, $(T(x), T(y))$ is the transformation value of pixel (x, y), and (± 1) denotes the direction of the image transformation. The effect of HWTDA is shown in Figure 1.

(**a**) (**b**)

Figure 1. Experimental results of Horizontal Wave Transformation. (**a**) The original image; (**b**) the image processed by HWTDA with R = 5, 15, 25, 35, and 45 from left to right.

(2) From the perspective of image appearance, an improved random erasing data augmentation algorithm suitable for Mongolian handwritten characters is proposed in this work. The random erasure algorithm is changed to selectively erase the pixels in the image within the specified range in order to retain more important information. In addition, according to the actual application scenario, the random erasure area is set as an oval. These modifications can improve the quality of the model training data to a certain extent, thereby improving the model's performance.

First, in the existing random erasing data augmentation algorithms, the content to be erased is determined by two random values: the randomly selected point $P = (x_t, y_t)$ in the image range and the random area ratio of the region to be erased to the whole image. For the object detection task, as long as the detection subject in the image is not blocked by a large area, the detectability of the image can be maintained, but the character recognition task has its own particularities, especially for the traditional Mongolian script. Since the writing format of letters is closely related to the location in traditional Mongolian script, if

the occlusion area generated by the random erasing data augmentation algorithm covers the main part of the image to be recognized, it is difficult to accurately distinguish the text, even manually. Therefore, the hasty use of random erasing data augmentation algorithm to augment Mongolian handwritten images may be not suitable for model training and may fail. Second, the stains that appear in handwritten characters are usually caused by ink, which is generally round or elliptical, but the random erasing data augmentation algorithm is originally applied to handwritten characters with square noise, which is not in line with reality.

This article has made the following improvements to overcome the above problems. First, the selection range of random points is specified to ensure that the occlusion caused by erasing will not affect the handwriting recognition of the image. Second, in order to make the occlusion area similar to the real ones and ensure the diversity of the random erasing area, the rectangle in the existing random erasing data augmentation algorithm is replaced by an ellipse. Through the dual data augmentation algorithm, more high-quality training data can be generated to improve the performance of the Mongolian handwriting recognition model.

The flow of the improved REDA algorithm is as follows:

Step 1: Enter an input image and initialize the algorithm parameters. Let the input image be I, with a size of S, and $S = W * H$, where W and H represent the width and height of image I, respectively. Set the erasing area ratio range $[S_l, S_h]$ and the erasing major to short axis ratio range $[r_a, r_b]$ and initialize the erasing probability p.

Step 2: Select an elliptical region with a random area. Let the area of the elliptical region I_t be S_t and the ratio of the major axis a to the minor axis b be r_t. They are randomly initialized to S_t and r_t, with S_t/S within range $[S_l, S_h]$ and r_t within range $[r_a, r_b]$. The formula for calculating the major and minor axes is as follows.

$$a = 2\sqrt{\frac{S_t r_t}{\pi}} \tag{3}$$

$$b = 2\sqrt{\frac{S_t}{\pi r_t}} \tag{4}$$

Step 3: Determine the random erasing location. A location point $P = (x_t, y_t)$ in the image I is randomly generated and it satisfies the requirements $\frac{a}{2} \leq x_t \leq \alpha W$ or $(1-\alpha)W \leq x_t \leq W - \frac{a}{2}$, $\frac{b}{2} \leq y_t \leq H - \frac{b}{2}$, with $0 < \alpha < 0.5$. A parameter of α is used to ensure that the occlusion caused by erasing will not affect the recognition of Mongolian characters in the image.

Step 4: For the selected elliptical area I_t, erase its pixels according to a probabilistic random value p.

The performance of the improved REDA algorithm is shown in Figure 2.

(**a**) (**b**) (**c**)

Figure 2. Experimental test of Improved Random Erasing Data Augmentation. (**a**) The original image; (**b**) the original REDA algorithm processes images through rectangular areas; (**c**) the improved REDA algorithm processes images through elliptical areas.

This paper proposed an improved dual data augmentation (DDA) algorithm by combining the above two data augmentation methods, and its flow is presented in Figure 3.

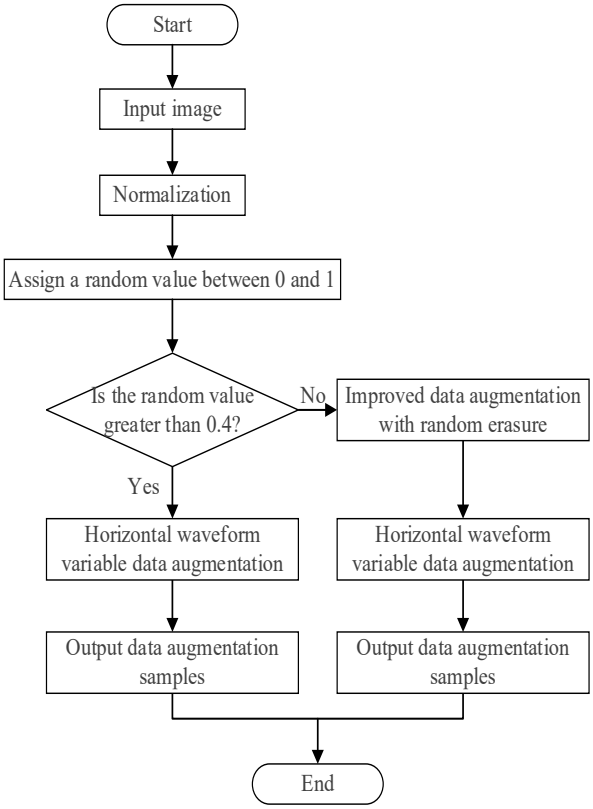


Figure 3. Flow of the Dual Data Augmentation Algorithm.

As shown in Figure 3, first, the Mongolian handwriting data are input into the model and then normalized. Second, a random value within [0, 1] is assigned by the system, and if the random value is greater than 0.4, only use the HWTDA method and output data; otherwise, the DDA method is adopted, that is, the input image data are randomly erased and then, the HWTDA is performed to output the final data augmentation result. In this method, data augmentation of an image is determined by a random value, which alleviates the problem that the similarity between the randomly erased image and the original image is too high. The method of random erasing first and then using horizontal wave transform can also give richer image occlusion noise, which helps improve data richness.

3. Improved Mongolian Handwriting Recognition Model: EGA Model

CRNN can realize end-to-end character recognition and is a landmark model in the development of character recognition. However, a large number of experimental results have shown that CRNN has low recognition accuracy when performing Mongolian handwriting recognition tasks; therefore, it is difficult to be used practically. So, this paper presents an ECA-Net-GRU-Attn (EGA) Model based on the characteristics of Mongolian handwriting.

3.1. Training Process of the EGA Model

Since the proposed EGA model is constructed by improving the CRNN, it follows the CRNN network structure and consists of an ECA-Net for feature extraction, a BiGRU encoder and a BiLSTM decoder for processing feature sequences, and an attention module to enhance sequence prediction performance.

The processing flow of an input image in the EGA model is shown in Figure 4.

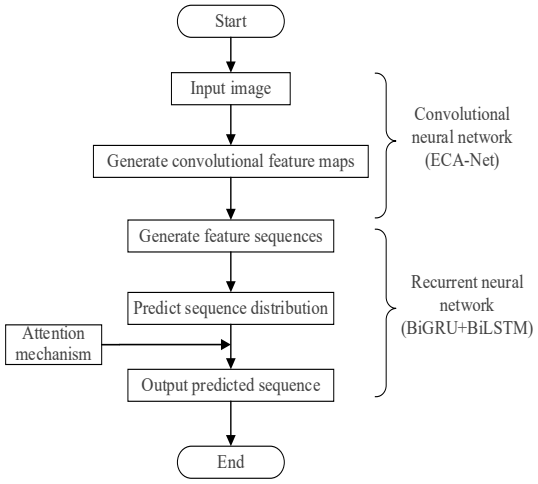
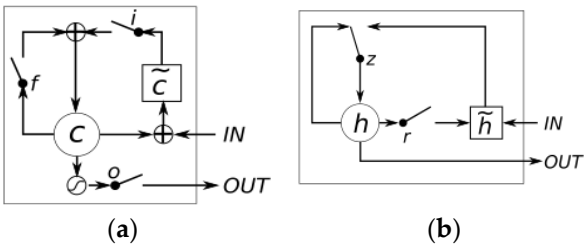


Figure 4. EGA Model Training Process.

First, compared to English handwritten script, the structure of traditional Mongolian handwritten script is more complex and flexible, which means that the feature extraction of images is more essential and difficult. The channel attention mechanism introduced in the feature extraction stage can grasp more detailed features of Mongolian handwriting images in a more directional way, improving the recognition accuracy of the model in recognizing Mongolian handwriting. In the EGA model, an ECA module, which is a channel attention module without dimensionality reduction, is utilized in the feature extraction stage of character recognition, which can overcome the shortcomings of insufficient feature extraction and insufficient processing of key areas. The ECA-Net has good experimental results in large-scale image classification, object detection, and instance ImageNet and MS segmentation [23], but we try to apply the technique in character recognition for the first time. Considering the complexity of the model, an ECA-Net model is designed by combining the ECA module with ResNet34 to extract the feature of Mongolian handwritten images.

Second, to balance the relationship between model recognition accuracy and complexity, the EGA model uses the BiGRU encoder instead of the BiLSTM encoder in the CRNN baseline model, which helps to reduce the overall complexity of the model. GRU and LSTM belong to recurrent neural networks based on the working mechanism of gating, and GRU is an evolved version of LSTM, reducing the original three gating to two. The diagrams of their structures are shown in Figures 5 and 6, respectively.



(**a**) (**b**)

Figure 5. Network LSTM and Network GRU Structure Comparison. (**a**) LSTM Network Structure; (**b**) GRU Network Structure.

From the comparison, it is clear that the structure of the GRU is significantly simpler. In addition, the EGA model performs the task of single-word recognition, and the training image used is homemade standard Mongolian handwriting recognition images by our laboratory; although data augmentation is added, the overall complexity of images is still low; therefore, GRU is fully qualified for encoding. The decoder follows the BiLSTM structure in CRNN, which has a more powerful ability to contact the context, helping to ensure the decoding quality.

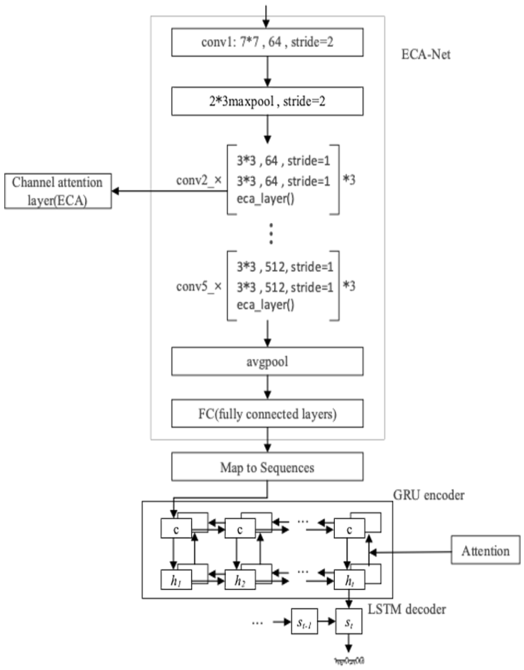


Figure 6. Network Structure of the EGA Model.

Finally, unlike English, Cyrillic Mongolian and other scripts, such as traditional Mongolian script, are in the form of a conjunction of letters from top to bottom and the beginning, middle, and ending writing formats of each letter in words are not exactly the same. Therefore, whether the context relationship of the training sample can be fully learned is crucial for the performance of traditional Mongolian handwriting recognition. The EGA model incorporates the attention mechanism into the encoder and decoder modules, which has good contextual analysis ability; therefore, introducing the attention mechanism in the output sequence prediction process can achieve further improvement of the CRNN model, which is helpful to increase the recognition accuracy of the model for Mongolian handwritten images.

3.2. Network Structure of the EGA Model

Process analysis of the EGA model can provide a more comprehensive understanding of the design concept of the model, and structural analysis of the network can deeply reveal the working principle of the model. The network structure of the EGA model is presented in Figure 6, and the way information is transmitted by each module of the EGA model can be seen more clearly through the network structure.

As shown in Figure 6, the ECA-Net is used to extract the input image features in the EGA model and the backbone network of the ECA-Net is the ResNet34, which consists of 33 convolutional layers as well as one maximum pool layer. Except for the first convolutional layer, every two convolutional layers form a residual unit and the Relu() function is executed between the residual elements as an activation function. The Mongolian handwritten input image is rotated 90 degrees to the left to form a rectangular structure with an aspect ratio of 75:40. The sliding window size of the pool layer in the ECA-Net is set to 2×3 to better collect image information.

The ECA module, which is a channel attention mechanism, can improve the model's attention to important channel features, thereby improving the generalization ability of the model on training and testing data. By considering the weights of different channels, the model can better capture the important features of Mongolian handwritten characters and improve the model's performance. Specifically, the channel attention mechanism can collect the interaction information between each channel and its k adjacent channels in a non-dimensionality-reducing manner and calculate the weight of the channel. In

the subsequent training process, the channel attention weight parameters learned by the module are represented by a matrix W_k and the expression of W_k is given in (5):

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{5}$$

The matrix W_k consists of $k * C$ parameters and C represents the size of the input image feature matrix, i.e., the number of output channels passing through the previous residual unit. The weight calculation formula for the image channel y_i is shown in (6):

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w_i^j y_i^j\right),\ y_i^j \in \Omega_i^k \tag{6}$$

where y_i^j represents the j-th neighboring channel of y_i, w_i^j represents the weight of y_i^j, and Ω_i^k represents the set of k adjacent channels of y_i. The dynamic value of k is proportional to C, as shown in (7):

$$C = \phi(k) = 2^{(\gamma+k-b)} \tag{7}$$

The values for C and k are adaptively adjusted according to Equation (8):

$$k = \psi(C)\left|\frac{log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{8}$$

where $|x|_{odd}$ represents the nearest odd number of x, and γ and b are set to be fixed constants. Due to the mapping function ψ, high-dimensional channels have a longer range of interactions, while low-dimensional channels experience a shorter range of interactions.

The ECA-Net adopts a global average pool layer at the end to connect a fully connected layer, passes the extracted Mongolian handwritten image feature map to the Map-to-Sequence layer through this fully connected layer, and converts it into a feature sequence. The feature vectors in the feature sequence are passed to the RNN as encoder input values.

The proposed EGA model has an encoder-decoder structure in the RNN. The configuration of the BiGRU encoder and the BiLSTM decoder can balance the complexity and the recognition accuracy of the Mongolian handwriting recognition model. Like the baseline CRNN model, the function of the cyclic layer in the EGA model is to predict the label distribution of each feature vector in the feature sequences, and the errors of the loop layer are backpropagated, converted into a feature sequence, and then fed back to the convolutional layer.

The input content of the GRU contains the input of the current timestep x^t and the hidden state h^{t-1} of the output of the previous timestep, and the hidden state contains the relevant information of the previous node. With x^t and h^{t-1}, the GRU can calculate the model output for the current timestep and the hidden state that will be passed to the next timestep. Specifically, this is achieved through the following steps:

(1) The gating status of the reset gate and the update gate are obtained by using h^{t-1} and x^t and are then normalized through the sigmoid function so that it acts as a gating signal. The operation of the reset gate in GRU is given in Equation (9):

$$r = \sigma\left(W_r x^t + U_r h^{t-1}\right) \tag{9}$$

where σ represents the sigmoid function and W_r and U_r are the weights learned in the reset gate. The calculation method for the update gate is shown in Equation (10):

$$z = \sigma\left(W_z x^t + U_z h^{t-1}\right) \tag{10}$$

W_z and U_z are the weights learned in the update gate.

(2) The hidden state of the current timestep t is calculated by Equation (11):

$$h^t = zh^{t-1} + (1-z)\tilde{h}^t \tag{11}$$

where $\tilde{h}^t$ contains the input of the current timestep and at the same time, the hidden state of the hidden unit of the previous timestep is added into the current hidden state; the calculation formula is shown in (12):

$$\tilde{h}^t = \phi\left(Wx^t + U\left(r \bigodot h^{t-1}\right)\right) \tag{12}$$

where W represents the weight of the current input and U represents the weight of the hidden state of the hidden unit in the previous timestep.

In the GRU, when the update gate approaches 0, the hidden state ignores the previous hidden state and updates only with the current input, which ensures that the hidden state selectively retains relevant important information and that the information can be expressed more compactly.

In summary, in the model training process, the hidden state h^{t-1} passed down from the previous timestep and the input x^t of the current timestep are used as the input of BiGRU. On one hand, it obtains the two gating states of the reset gate and the update gate and normalizes the obtained information through the sigmoid function to make it act as a gating signal. On the other hand, the image feature sequence is converted into a context vector C_t as the output.

This work introduces an attention mechanism to the encoder-decoder structure to give weights to the output of the encoder. The main implementation details of the attention mechanism are as follows: first, the input data are fed through a feedforward neural network to generate query vectors and key vectors. Then, the similarity between the query vectors and key vectors is calculated and the resulting scores are normalized to obtain the final attention weights. Finally, the attention weights are applied to the value vectors and summed up to obtain the weighted sum as the output feature vector. Through this process, the attention mechanism can learn the importance of different parts of the input data and enable the decoder to better utilize the information from the encoder. It also has strong interpretability and generalization ability.

Specifically, the attention mechanism learns a special attention weight by adding an additional feedforward neural network to the network structure α_{t_e,t_d}, where t_e and t_d represents the t-th timestep of the encoder and decoder, respectively. The introduction of this weight into the neural network helps to further reconcile the relationship between the hidden states in the encoder and decoder to highlight the focus of model training. The context vector C_t of the output of the t-th timestep encoder can be assigned by Equation (13):

$$C^t = \sum_{t_e=1}^{T} \alpha_{t_e,t_d} h_{t_e} \tag{13}$$

At the t-th timestep of the decoder, C_t is the weighted sum containing all encoder hidden states and their corresponding attention weights.

The structure of the decoder BiLSTM is similar to that of the encoder. When calculating the prediction sequence, the input values accepted by the decoder are the context vector C_t weighted by the attention mechanism, the system state S_{t-1} of the output of the decoder at the previous timestep, and the hidden state S_{t-1} of the decoder at the previous timestep. With the help of such information, the decoder can calculate the hidden state h_t of the current time node and the probability distribution of the output sequence of the node according to the probability calculation formula, as shown in (14).

$$p(S_t|S_{t-1}, S_{t-2}, \cdots, S_1, C_t) = g(h_t, S_{t-1}, C_t) \tag{14}$$

where g represents the given activation function. The softmax is used as the activation function for this stage in the EGA model. The largest output value of $p(S_t|S_{t-1}, S_{t-2}, \cdots, S_1, C_t)$ is treated as the output value of the t-th timestep. Ultimately, the output values of each timestep make up the output sequence of the model.

4. Data Experimental

4.1. Experimental Data

The dataset used in this article comes from the Inner Mongolia Normal University and includes 9479 handwritten Mongolian characters, totalling 47,395 images. The dataset covers various handwriting styles and samples from different writers, with a wide variety of fonts, and each character has a clear shape, which is highly representative and can truly reflect the features of Mongolian handwriting styles. Furthermore, the dataset is balanced, i.e., the number of images per character is similar among them. All these characters are carefully handwritten and drawn by writers, closely related to practical application scenarios and are highly authentic and practical, making them suitable for data training and testing for handwriting character recognition tasks. Some offline Mongolian handwriting sample images of the dataset are shown in Figure 7.

Figure 7. Offline Mongolian Handwriting Images.

4.2. Model Training Platform

According to the training requirements of the Mongolian handwriting recognition model, the model training environment built in this paper is shown in Table 1.

Table 1. Model Training Platform Parameters.

Name	Parameters
CPU	Intel Core i7-6700 CPU@3.40 GHz
GPU	Nvidia Tesla P100 + Huawei GPU Server
Operating system	Ubuntu 16.04.6
Programming language	Python 3.7
Deep learning framework	Pytorch 1.9.0

4.3. Indicators for Model Performance Evaluation

The quality of the Mongolian handwriting recognition model is primarily measured during the training process from two aspects: the recognition accuracy for the test set and the complexity of the trained model.

(1) Accuracy is the primary criterion to measure the effectiveness and practicality of Mongolian handwriting recognition models and it can be calculated by the confusion matrix, in which the prediction accuracy of the classification algorithm is evaluated by the relationship between the true category of the sample and the predicted value of the model, as shown in Table 2.

Table 2. Confusion Matrix.

	Forecast Category	
Real category	Positive category	Negative category
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

In the confusion matrix, TP and TN represent the situation where the predicted category matches the real category and the classification if the predicted value is correct; FP and FN represent the situation where the predicted category does not match the real category and the classification is wrong. The accuracy is defined by the proportion of correctly classified data to the total test set and calculated by Equation (15).

$$Accurary = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

In Mongolian handwriting recognition, predicted value classification is more diverse and not limited to positive and negative categories, and the correct classification represents the situation where the sequence output of the model is consistent with the sequence of image labels.

(2) Model complexity is another important criterion for measuring model quality. Under the condition of the same test platform and data, the model training time, the number of model parameters generated during the training process, and the size of the trained model are used as indicators to evaluate the model's complexity in this work.

4.4. Data Test of the Dual Data Augmentation Method and Result Analysis

The dual data augmentation algorithm combines random erasing with horizontal wave transformation, and its parameters also need to be uniformly initialized. The test parameters of dual data augmentation are listed in Table 3. The optimal hyperparameter values mentioned above were used to ensure the sample quality of the proposed data augmentation.

Table 3. Dual Data Augmentation Hyperparameter.

Hyperparameter Name	**Hyperparameter Meaning**	**Hyperparameter Value**
p	A random value that determines the image data augmentation method	[0, 1]
W	Image width	800
H	Image height	1500
S_l	For random erasing, the minimal erasing area ratio	0.01
S_h	For random erasing, the maximal erasing area ratio	0.06
r_a	The minimal ratio of the major to minor axes of the erasing area	0.5
r_b	The maximal ratio of the major to minor axes of the erasing area	2

The effect of using dual data augmentation to achieve data augmentation is shown in Figure 8. The images are, from left to right, the original image, the image processed only by HWTDA, and the image processed by DDA.

(**a**) (**b**) (c)

Figure 8. Effect of Dual Data Augmentation. (**a**) The original image; (**b**) the image processed only by HWTDA; (**c**) the image processed by DDA.

After DDA, the Mongolian handwriting recognition data have been effectively augmented, and the data before and after the augmentation are evenly mixed and randomly distributed to the training set and test set with a number ratio of 7:3 to form datasets for Mongolian handwriting recognition. The data augmentation and dataset segmentation results are summarized in Table 4.

Table 4. Dataset Segmentation after Dual Data Augmentation.

Data Type	Raw Data (Wordage × Frame Number)	Dual Data Augmentation (Wordage × Frame Number)		Total Data (Wordage × Frame Number)	Unit
		Horizontal Wave Transformation	Horizontal Wave Transformation + Random Erasing		
Training set	33,175 (6635 × 5)	33,175 (6635 × 5)	13,512 (6635 × 2)	79,862 (6635 × 12)	sheets
Test set	14,220 (2844 × 5)	14,220 (2844 × 5)	5791 (2844 × 2)	34,231 (2844 × 12)	sheets
Total	47,395 (9479 × 5)	47,395 (9479 × 5)	19,303 (9479 × 2)	114,093 (9479 × 12)	sheets

The direct purpose of designing and using dual data augmentation is to produce Mongolian handwritten images with more diverse forms and richer content, but the fundamental purpose is to deepen the understanding of Mongolian handwriting for the neural network model and improve the accuracy of Mongolian handwriting recognition through continuous learning of a large amount of data.

As shown in Figure 9, the CRNN model is trained separately by using the dataset before and after data augmentation, and the impact and significance of data augmentation can be better highlighted by comparing the training accuracy of the model. In Figure 9, the coordinate horizontal axis represents the model training epochs. The vertical axis represents the recognition accuracy (%) of the test set by the model during training.

From the accuracy curve, it can be seen that after 15 epochs of training, the recognition accuracy of the CRNN model trained using the raw dataset gradually converged to 56.2%. The growth rate of recognition accuracy of the CRNN models using data-augmentation datasets gradually slowed down but still showed an upward trend, and the recognition accuracy rate was 69.3%. The experimental results showed that DDA can help improve the recognition accuracy of the Mongolian handwriting recognition model.

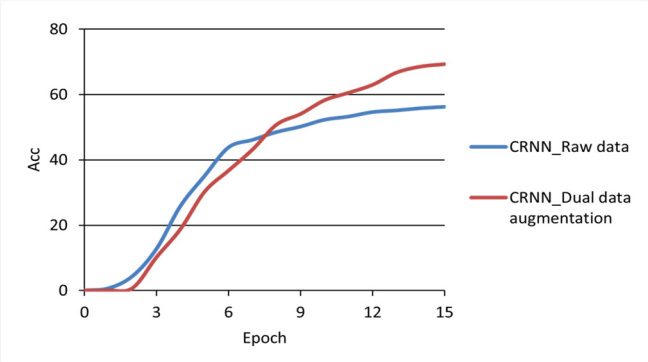


Figure 9. CRNN Model Recognition Accuracy Curve before and after Dual Data Augmentation.

This is also confirmed by the loss value curve of the model. Based on the loss values of each epoch for the test set, the model loss function curves are plotted in Figure 10. In the figure, the horizontal axis represents the model training epochs and the vertical axis represents the loss values of the model for the test set.

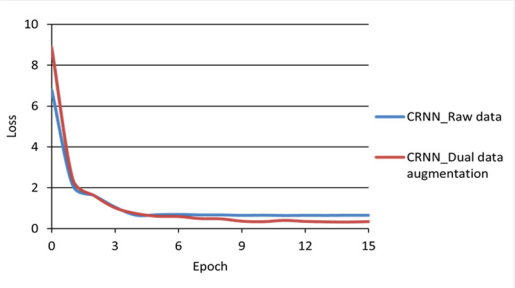


Figure 10. Model Loss Value Curve before and after Dual Data Augmentation.

It can be seen from the change trend of the loss value curve that after the model was trained for 15 epochs, the convergence speed of the CRNN model trained with the raw data and the dual data augmentation dataset gradually slowed down and the loss function converged to a lower value in the model with the dual data augmentation dataset, indicating that the dataset is more adequate for model training. The experimental results showed that the dual data augmentation effectively augmented the model training dataset and improved the recognition accuracy of the Mongolian handwriting recognition model.

4.5. Data Test for the EGA Model and Results

To ensure the training quality of the EGA model, we use the dataset augmented by DDA for model training.

The curves of the training loss value and the test loss value of the EGA model are presented in Figure 11, and the horizontal axis represents the model training epoch. The vertical axis represents the model loss values.

It can be seen that, with the increase in the number of iterations, the loss value of the training set and the test set of the model gradually decreased, and when the iteration reached about 25 epochs, the decrease in the loss value of the training set and the test set slowed down and gradually stabilized and the loss function of the test set converged to about 0.3. To prevent the model from being overfitted due to too many iterations, the training of the model was stopped.

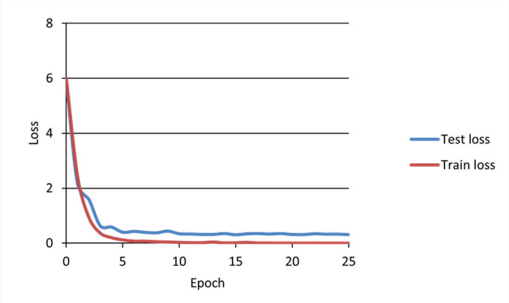


Figure 11. EGA Model Loss Value Curve.

In order to verify the improvement performance of the model, the EGA model of the experimental group and the CRNN baseline model of the control group were trained in the same experimental platform and parameter settings. Table 5 compares the performance of the EGA model and the CRNN model from the perspective of recognition accuracy and model complexity. The unit of the model scale is MB, the unit of the number of model parameters is M, the unit of accuracy is %, and the unit of the model training time is s.

Table 5. Performance Comparison Between EGA Model And CRNN.

Model	Model Size (MB)	Parameter Number (M)	Accuracy (%)	Time (s)
ECA-Net-GRU-Attn (EGA)	188.6	49.3154	89.322	215,112.2
CRNN	39.8	10.2280	74.307	196,519.8

With the recognition accuracy (%) of the model test set on the vertical axis and the model training epochs on the horizontal axis, the recognition accuracy curves of the EGA model and the CRNN model are shown in Figure 12. The recognition accuracy of the EGA model in the Mongolian handwriting recognition test set was better than that of the CRNN model, achieving the expected effect of the experiment.

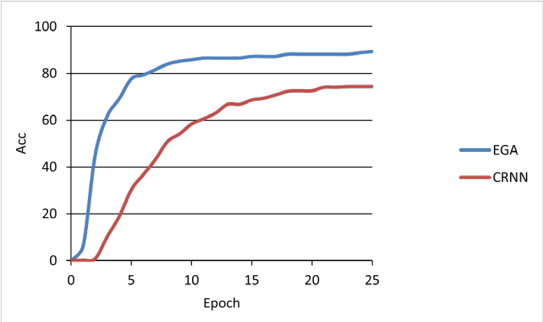


Figure 12. Comparison of Recognition Accuracy between EGA Model and CRNN Model.

From the perspective of model complexity, the size of the CRNN model and the number of parameters were small, the size of the EGA model was about 4.7 times that of the CRNN model, and the amount of parameters was about 4.8 times that of the CRNN model. In terms of training time, the training time of the CRNN model and the EGA model was 196,519 s and 215,112 s, respectively, indicating an increase of 18,592.4 s for the training time of the EGA model. It can be seen that with the increase in model depth, the model training efficiency decreases.

4.6. Comparison Test of Different Models and Result Analysis

In order to measure the impact of each module in the EGA model on performance and to demonstrate the advantages of the EGA model, comparative experiments for important modules of ECA, GRU, and Attn were conducted to test the recognition rate and training time of each model on the Mongolian handwriting dataset. The experimental results are summarized in Table 6.

Table 6. Comparison of Experimental Results of Each Model.

Model	Model Size (MB)	Parameter Number (M)	Accuracy (%)	Time (s)
ECA-Net-GRU-Attn (EGA)	188.6	49.3154	89.322	215,112.2
ResNet-GRU-Attn (RGA)	186.8	49.3153	85.496	207,709.5
ECA-Net-LSTM-Attn (ELA)	189.4	49.9728	90.846	223,211.8
ECA-Net-GRU-CTC (EGC)	184.3	48.2845	84.303	218,604.7
CRNN	39.8	10.2280	74.307	196,519.8
ResNet-GRU-Featfusion (RGF-CRNN)	115.4	30.1	85.905	215,112.7
Featfusion-ViT (F-ViT)	97.6	25.4	87.417	159,604.5

As shown in Table 6, from the perspective of the impact of the ECA module on the model recognition efficiency, the EGA model has a large scale, a large number of parameters, and a high recognition accuracy, with slightly longer model training time. Compared to the RGA model, the accuracy of Mongolian handwriting recognition was improved by 3.826%, and the model training time was increased by 7,402.7 s for the EGA model, indicating that, under the same experimental conditions, the introduction of the ECA module can improve the accuracy of handwriting recognition to a certain extent and prolong the model training time.

From the perspective of the impact of the GRU module on model recognition efficiency, compared to the ELA model, the model size and parameter amount are small, the Mongolian handwriting recognition accuracy was reduced by 1.524%, and the training time was increased by 8,099.6 s, indicating that the simplified GRU encoder module can reduce the complexity and training time of the model to a certain extent.

From the perspective of the impact of the Attn module on model recognition efficiency, compared to the EGA and EGC models, the EGA model has a larger scale and more parameters, the recognition accuracy was increased by 5.019%, and the training time was reduced by 3,492.5 s, showing that the introduction of the Attn module can improve the overall recognition efficiency of the model.

Furthermore, the experimental results in Table 6 demonstrate that the EGA model has certain advantages over three other models—CRNN, RGF-CRNN, and F-ViT—in the Mongolian handwritten characters recognition task. On the test set, the EGA model achieved the highest accuracy (89.322%) and showed significant improvement compared to the other three models. This indicates that the EGA model is more efficient at capturing the features of Mongolian characters.

In summary, introducing the ECA, GRU, and Attn modules can improve the recognition efficiency of the model to varying degrees, validating the effectiveness of the EGA model. From the perspectives of model size, parameters, accuracy, and training time, the EGA model has better recognition performance than the baseline model and performs well on data from different handwriting styles and writers, which is the main advantage of the model. However, the limitation of the model lies in its relatively larger parameter size and longer training time, which may occupy more storage space and consume more computing resources.

5. Conclusions

In this paper, Mongolian handwriting data were used as the research object, two methods were adopted to augment the Mongolian handwriting dataset, the CRNN model

and training set were used to establish a Mongolian handwriting recognition model, and the tests were conducted to evaluate the performance of the proposed model. The experimental results demonstrated that the overall accuracy of the CRNN handwriting recognition model was about 74.307% and the model training time was 196,519.8 s. In order to further improve the recognition performance of the CRNN model, we proposed an EGA model, which uses ECA-Net to realize feature extraction, a BiGRU encoder and BiLSTM decoder to process feature sequences, and an attention module to improve sequence prediction performance. The experimental results of each model were compared and analyzed on the augmented dataset. The EGA model can effectively improve the accuracy of Mongolian handwriting recognition. The size of the obtained EGA model was 188.6 MB, the number of parameters was 49.3154 M, the accuracy rate was 89.322%, and the model training time was 215,112.2 s in Mongolian handwriting recognition.

Although the performance of the Mongolian handwritten characters recognition model can be improved to some extent by using data augmentation methods and improving the network structure, the model has certain limitations due to restrictions in dataset size and structure. For example, the model has inadequate robustness for erroneous or exceptional characters and insufficient adaptability for multi-task or multi-scenario applications, and the prediction accuracy and the model performance still need to be further improved on larger datasets. In the future, we will optimize the EGA model based on the following two aspects: obtaining more data by collecting more datasets, such as introducing various data augmentation methods or considering the combined use of GANs and other generative models to generate more diverse and realistic Mongolian handwritten data, which can strengthen model training and improve recognition accuracy. In terms of model improvement, multi-task learning can be used, or more efficient feature fusion and model combination methods or training on large datasets can be adopted to enhance the model's performance and generalization ability. Using deep neural network structure to mine more valuable data information and shortening training time while maintaining the training effect is the focus of future research.

Author Contributions: Conceptualization, methodology, software, writing—original draft, Q.-D.-E.-J.R.; software, validation, writing—review, L.W.; software, writing—review, Z.M.; validation, writing—review and editing, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61966027), the Project of "Support Program for Young Scientific and Technological Talents" in Inner Mongolia Colleges and Universities (NJYT23059), Inner Mongolia Natural Science Foundation (2022MS06013), Inner Mongolia Science and Technology Program Project (2021GG0140), and Universities Directly Under the Autonomous Region funded by the Fundamental Research Fund Project (JY20220122).

Data Availability Statement: Data are contained within this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
2. Ciresan, D.C.; Meier, U.; Schmidhuber, J. Transfer learning for Latin and Chinese characters with deep neural networks. In Proceedings of the 2012 International Joint Conference on Neural Networks, Brisbane, Australia, 10–15 June 2012; pp. 1–6.
3. Shi, B.; Bai, X.; Yao, C. 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.* **2016**, *39*, 2298–2304. [CrossRef] [PubMed]
4. Tang, J.; Silamu, W.; Xu, M.; Xu, M.-M.; Xiong, L.-J.; Wang, M.-H. Uyghur scanning body recognition based on deep learning. *J. Northeast Norm. Univ. (Nat. Sci. Ed.)* **2021**, *53*, 71–76. [CrossRef]
5. Wang, D. Research and Application of Yi Online Handwriting Recognition. Master's Thesis, Southwest University, Chongqing, China, 2021.
6. Li, J. Tibetan-Chinese Bilingual Natural Scene Text Detection and Recognition System. Master's Thesis, Northwest University for Nationalities, Lanzhou, China, 2021.
7. Du, Y.; Li, C.; Guo, R.; Yin, X.; Liu, W.; Zhou, J.; Bai, Y.; Yu, Z.; Yang, Y.; Dang, Q.; et al. PP-OCR: A Practical Ultra Lightweight OCR System. *arXiv* **2020**, arXiv:2009.09941v3. [CrossRef]

8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]
9. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
10. Riaz, N.; Arbab, H.; Maqsood, A.; Nasir, K.; Ul-Hasan, A.; Shafait, F. Conv-transformer architecture for unconstrained off-line Urdu handwriting recognition. *Int. J. Doc. Anal. Recognit.* **2022**, *25*, 373–384. [CrossRef]
11. Yongping, D.; Zhuo, L. Particle Swarm Optimization-Based Convolutional Neural Network for Handwritten Chinese Character Recognition. *J. Adv. Comput. Intell. Intell. Inform.* **2023**, *27*, 165–172.
12. Fan, D.; Gao, G.; Wu, H. Research on Mongolian handwriting recognition based on morpheme segmentation. *J. Chin. Inf. Technol.* **2017**, *31*, 74–80.
13. Liu, C. Research on Recognition of Large Vocabulary Off-Line Handwritten Mongolian Whole Words. Master's Thesis, Inner Mongolia University, Hohhot, China, 2019.
14. Wei, H.; Zhang, H.; Zhang, J.; Liu, K. Multi-Task Learning Based Traditional Mongolian Words Recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1275–1281.
15. Yang, F. Research and Implementation of Mongolian Online Handwriting Recognition Based on Whole Word. Master's Thesis, Inner Mongolia University, Hohhot, China, 2021.
16. Luo, C.; Zhu, Y.; Jin, L.; Wang, Y. Learn to augment: Joint data augmentation and network optimization for text recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13743–13752.
17. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 34.
18. Han, Y. Research on Data Enhanced Ancient Chinese Character Recognition Method. Master's Thesis, Xiamen University of Technology, Xiamen, China, 2021.
19. ul Sehr Zia, N.; Naeem, M.F.; Raza, S.M.K.; Khan, M.M.; Ul-Hasan, A.; Shafait, F. A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition. *Neural Comput. Appl.* **2021**, *34*, 1635–1648. [CrossRef]
20. Eltay, M.; Zidouri, A.; Ahmad, I.; Elarian, Y. Generative adversarial network based adaptive data augmentation for handwritten Arabic text recognition. *PeerJ Comput. Sci.* **2022**, *8*, e861. [CrossRef] [PubMed]
21. Maalej, R.; Kherallah, M. New MDLSTM-based designs with data augmentation for offline Arabic handwriting recognition. *Multimed. Tools Appl.* **2022**, *81*, 10243–10260. [CrossRef]
22. Gao, H.; Ergu, D.; Cai, Y.; Liu, F.; Ma, B. A robust cross-ethnic digital handwriting recognition method based on deep learning. *Procedia Comput. Sci.* **2022**, *199*, 749–756. [CrossRef]
23. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. Eca-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11531–11539.