

Article

Large-Scale Printed Chinese Character Recognition for ID Cards Using Deep Learning and Few Samples Transfer Learning

Yi-Quan Li ^{1,2} , Hao-Sen Chang ¹  and Daw-Tung Lin ^{1,*} 

¹ Department of Computer Science and Information Engineering, National Taipei University, 151, University Rd., San-Shia, New Taipei City, Taiwan

² Orbit Technology Inc., 5F, No. 126, Minzhu West Road, Datong District, 10342, Taipei City, Taiwan

* Correspondence: dalton@mail.ntpu.edu.tw

Abstract: In the field of computer vision, large-scale image classification tasks are both important and highly challenging. With the ongoing advances in deep learning and optical character recognition (OCR) technologies, neural networks designed to perform large-scale classification play an essential role in facilitating OCR systems. In this study, we developed an automatic OCR system designed to identify up to 13,070 large-scale printed Chinese characters by using deep learning neural networks and fine-tuning techniques. The proposed framework comprises four components, including training dataset synthesis and background simulation, image preprocessing and data augmentation, the process of training the model, and transfer learning. The training data synthesis procedure is composed of a character font generation step and a background simulation process. Three background models are proposed to simulate the factors of the background noise and anti-counterfeiting patterns on ID cards. To expand the diversity of the synthesized training dataset, rotation and zooming data augmentation are applied. A massive dataset comprising more than 19.6 million images was thus created to accommodate the variations in the input images and improve the learning capacity of the CNN model. Subsequently, we modified the GoogLeNet neural architecture by replacing the FC layer with a global average pooling layer to avoid overfitting caused by a massive amount of training data. Consequently, the number of model parameters was reduced. Finally, we employed the transfer learning technique to further refine the CNN model using a small number of real data samples. Experimental results show that the overall recognition performance of the proposed approach is significantly better than that of prior methods and thus demonstrate the effectiveness of proposed framework, which exhibited a recognition accuracy as high as 99.39% on the constructed real ID card dataset.

Keywords: Large-Scale Image Classification; Printed Chinese Character Recognition; Data Synthesis; GoogLeNet-GAP; Transfer Learning



Citation: Li, Y.-Q.; Chang, H.-S.; Lin, D.-T. Large-Scale Printed Chinese Character Recognition for ID Cards Using Deep Learning and Few Samples Transfer Learning. *Preprints* **2021**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Image classification has always been one of the prominent topics in deep learning, and Chinese character recognition is one application of it. Traditionally, optical character recognition (OCR) has been used for text recognition and it has achieved good results. Large-scale image classification is an important and challenging task in the field of computer vision, which plays an essential role in facilitating OCR methods. For example, the number of classes for a Chinese OCR system could be as high as 13,070. When performing large-scale classification, the amount of data in each category is considered the most important factor. By contrast, if a classifier is divided into an excessive number of characters, the accuracy decreases with an increase in the numbers of characters.

ID card information verification is widely performed for multiple purposes on various occasions, such as for opening bank accounts or making deposits, hotel check-in, clinic registration, identity verification at facility entrances, and pick-up services for purchased items. The development of a system designed to automatically identify personal data on ID cards is expected to provide considerable convenience for both customers and service providers. It also saves human resources and reduces the possibility of errors. It not only saves time but also reduces physical contact, especially when infectious diseases

are prevailing; it is especially important and safe. Hendra Dito Dwi et al. [1] developed an OCR system to identify new ID cards issued in Indonesia. Similarly, Angga Maulana et al. [2] used MSER to detect pre-processed Indonesian ID card images and found the area where the text was located. Wira et al. [3] applied a series of image processing techniques, such as image binarization, Sobel edge detection, and morphology, to mark the text area on citizen ID cards. Then, Google Tesseract was used as a primary framework for character recognition. Their approach correctly identified citizen ID cards at a rate of over 90%. Niloofar et al. [4] designed a network model called efficient and accurate scene text detector (EAST), which can accurately extract the text area. Compared with the MERS-based algorithm, it is more adaptable to natural noise and is faster.

On arriving at a hotel, conventionally, one must check-in via the reception staff. The traditional method involves manual data entry on a workstation computer through the manual confirmation of identity documents. Although a bar code is included in contemporary ID cards that records personal information, only government agencies or specific institutions can access and use it owing to privacy issues; ordinary hotels cannot use this feature. Human error is typical under such settings. For example, during the peak tourist season, the influx of a large number of customers may easily cause the reception staff to panic or otherwise perform imperfectly, leading to data entry errors. Moreover, customers’ privacy may be easily violated and their personal information may be exploited by the malicious actions of staff at the reception desk itself. In this study, we aim to establish a self-service check-in system, as shown in Figure 1, utilizing the proposed large-scale printed Chinese OCR with deep learning and transfer learning. When a user presents their ID card to the camera device, the system can automatically recognize their personal information and complete the follow-up check-in procedures, which not only solves the manual error problem but also avoids the possibility of malicious actions.

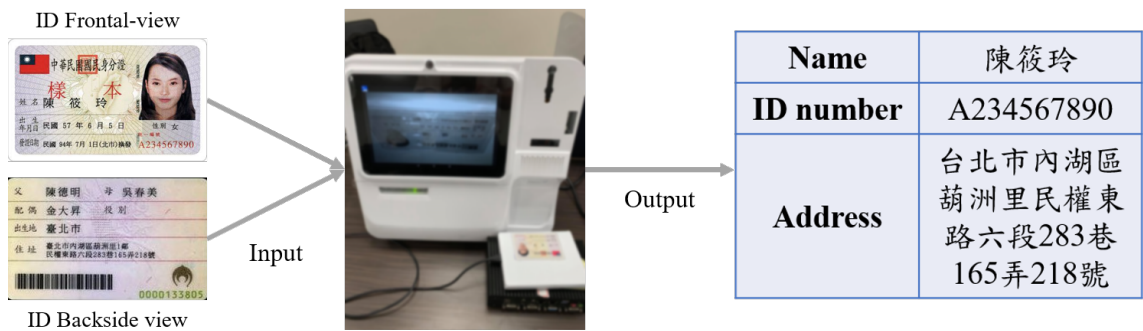

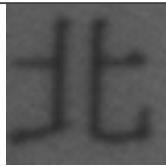


Figure 1. Prototype of the self-service check-in system.

1 Although handwritten text recognition methods have been established, further re-
2 search on text recognition on ID cards remains necessary to account for the variations in
3 backgrounds and lighting. In real situations, there will be different factors to be considered
4 such as lighting conditions, the material of the paper used, color, and text font will also have
5 different interference. Various background patterns and anti-counterfeiting mechanisms
6 are also being added to the ID cards. The influence of background noise tends to cause
7 problems in character segmentation and recognition. Furthermore, after using the ID cards
8 for a long period, they can have all kinds of scratches and damages, and the lamination
9 film of the cover can turn yellow with aging. The interweaving of these factors increases
10 the difficulty of identification. Another challenge in Chinese character recognition is a large
11 number of Chinese characters. Unlike digits or English alphabets, Chinese characters are
12 more than tens of thousands of characters, and even the commonly used Chinese characters
13 have more than 4000 categories. Moreover, most of the research focuses on handwritten
14 text, and few have classified or identified printed text. Handwritten text is deliberately
15 collected, which is mostly generated in a laboratory or a single environment. The text image
16 was clear, and the quality of the image was better. There is no need to perform excessive

preprocessing. Table 1 compares the characteristics of handwritten character images and our collected ID card character images. To strengthen the adaptability and improve the performance of the proposed ID card text recognition method, we implemented dataset synthesis, background simulation, and transfer learning.

Table 1. Comparison between handwritten text and our ID card text.

	Handwritten Text	Our Input Text
Sample Image "North"		
Complexity	High	High
Background	Clean	Noisy
Image Size	Specify	Small

These challenges motivated us to study the problem of a large-scale classification model for ID card character recognition. In this study, we developed an automatic OCR system to identify up to 13,070 large-scale printed Chinese characters using deep learning neural networks and fine-tuning techniques. As shown in Figure 2, our framework consists of four components. (1) Training dataset synthesis is used to generate a synthesized training set of character images with different Chinese character fonts and simulated ID card backgrounds. (2) Data augmentation is performed on the synthesized training dataset by applying rotation and zooming to increase the diversity of the images contained. (3) The synthesized training dataset is input to the modified GoogLeNet Inception network (GoogLeNet-GAP) [5] for training. (4) Finally, we collected several samples of real Chinese characters on ID cards and performed data augmentation and balancing processing for further transfer learning; finally, the recognition results were obtained as the output.

The main contributions of this work are summarized as follows.

- We developed a large-scale OCR system to identify printed Chinese characters using a deep learning neural network.
- We propose a new procedure for synthesizing printed Chinese characters that simulates the factors of background noise and anti-counterfeiting patterns on ID cards. A massive dataset of more than 19.6 million images was created to accommodate the diversity of input image variations and strengthen the recognition capability of the CNN model.
- We improve the recognition performance of the GoogLeNet-GAP model by incorporating transfer learning.
- The experimental results demonstrate the effectiveness of the proposed framework; the accuracy of the large-scale 13,070 character recognition system was as high as 99.39%, as evaluated on our dataset of images of real ID cards.

2. Related Work

2.1. Deep Learning Chinese Character Recognition

Optical character recognition (OCR) is a mainstream text recognition technology. Through the analysis and comparison of a pre-built database, various texts can be quickly recognized. With the tremendous advancement of deep learning-based computer vision technology in recent years, the convolutional neural network (CNN) approaches have been greatly applied to OCR. Xu et al. [6] proposed an end-to-end subtitle recognition system. After inputting a video with subtitles, the subtitle area is marked, and sliding windows are

used to cut the characters one by one at regular intervals and recognize them. Although an end-to-end system is adopted, the sliding window can easily cut out too many repetitive characters and cause failure. Zhong et al. [7] proposed an HCCR-GoogLeNet, which reduces the parameters of the original GoogLeNet [5], and adds the traditional feature extraction method HoG, and obtain gradient feature maps to enhance the performance of CNN. The recognition rate of the model evaluated with the CASIA-HWDB dataset [8] attained 96.74%. Lin et al. [9] proposed a new architecture to avoid the overfitting problem under a multi-classification problem. By using global average pooling (GAP) to replace the original fully connected (FC) layer, it can reduce the number of parameters while maintaining the original effect. Li et al. [10] resolved the disadvantage of the traditional CNN, which would require huge computation resources, and improved the recognition effect by increasing the output layer of the intermediate layer. However, GAP will cause a significant decrease in accuracy, hence they added trainable weights to GAP to solve this problem which is called global weighted average pooling (GWAP). The model achieved an accuracy of 97.1% for the CASIA-HWDB1.1 [8] dataset. Xiao et al. [11] proposed a new technology called adaptive drop weight (ADW), which can effectively reduce the number of CNN parameters, and proposed global supervised low-rank expansion (GSLRE) to accelerate the entire CNN model. Compared to other baselines of handwritten Chinese character recognition (HCCR) models, the model reduces the amount of calculation by nine times and compresses the parameter amount to 1/18, but the accuracy rate is only reduced by 0.21%. With its extremely low parameter amount, it can be deployed on a mobile device to achieve rapid recognition. Melnyk et al. [12] proposed a new architecture in 2019 to improve the classification accuracy of the output layer by modifying GWAP, and extracting the class activation map (CAM) to perform character recognition. This model was evaluated with the CASIA-HWDB1.1 dataset [8], and the recognition accuracy rate reached 97.61%, breaking the record of the year. Su et al. [13] developed a device that could be worn on the hand. This model can accurately identify 6,000 Chinese characters and achieves an accuracy of 96.75% in the Chinese FingerReader testing dataset. However, its research also shows that the overall recognition rate will be affected if it encounters serious background interference. Liu et al. [14] built a CNN model with connectionist temporal classification loss function. To reduce the overfitting, we applied dropout after each max-pooling layer. They achieved 6.81% character error rate on the ICDAR 2013 competition set.

2.2. Handwritten Chinese Character Dataset

Recent Chinese character recognition researches are mostly evaluated using the CASIA-HWDB dataset [8]. CASIA-HWDB [8] was released in the ICDAR 2013 offline HCCR competition, which contains three subsets, CASIA-HWDB1.0, 1.1, and 1.2. Generally subsets 1.0, and 1.2 are used as the training set, and subset 1.1 is used as the test set, which contains 3755 different handwritten Chinese characters. However, these datasets are simplified Chinese and not traditional Chinese. Traditional Chinese and simplified Chinese are quite different in structure. The simplified Chinese omits complicated strokes for the convenience of writing. Therefore, the recognition of simplified Chinese is generally easier than that of traditional Chinese. There are not many traditional Chinese datasets. Among them, Chinese MNIST [15] is the largest handwritten traditional Chinese dataset, which contains 13,065 traditional Chinese characters. Although the number of characters is much larger than that of CASIA-HWDB [8], the amount of samples of each character is far behind. If CASIA-HWDB is used for a large-scale classification training, it is very possible to cause underfitting because of insufficient data. Yue et al. [16] established a database CASIA-AHCDB of ancient characters. It was collected from 11,937 pages of ancient Chinese manuscripts. There are 10,350 characters of different handwritten Chinese. They are mainly divided into two large datasets, each containing three parts of the data, which can be used for different purposes. However, the number of characters differs. If they are used as the

training data for a large classifier, the convergence time will be too long or even unable to converge.

2.3. Large-Scale Classification

When performing large-scale classification, the most important factor is the amount of data in each category. To train a high-precision neural network, a large amount of data is necessary to assist, and it is known as data hungry. If the data are insufficient, the model will not be able to converge, or the amount of data in each category will be different, which will easily bias the model to the side with more data. Therefore, in the research conducted so far, there are few relevant interpretations for large-scale classification. Once the classifier is divided into too many characters, the accuracy decreases as the number of characters increases. Although the impact can be reduced by data expansion, it still cannot reach a practical level. The traditional neural network model uses the fully connected (FC) architecture in the final output layer to classify the features extracted in the first half of the CNN. However, because of its too many parameters, it not only increases the burden on calculations but also easily causes overfitting problems. Therefore, partial connections are often used to reduce the probability of overfitting. If it is used for Chinese character classification, the impact will be even more serious.

Zhong et al. [17] proposed a multi-pooling method and nonlinear data transformation to improve the effect of large-scale printed Chinese character recognition. Experimental results show that the proposed model achieves good results for 3,755 printed Chinese characters. Li et al. [10], Melnyk et al. [12], and Qiu [18] demonstrated through experimental results that GWAP with adaptable parameters can extract regional features better than traditional fully connected layers to improve the classification effect and reduce the burden on calculations. Zhang et al. [19] proposed a label-mapping (LM) strategy. By dividing a huge category into several sub-characters and then predicting each sub-category, the results showed that LM can effectively improve the accuracy of large-scale classification on the CJK characters. Zhang et al. [20,21] proposed the radical analysis network (RAN), which analyzes the two-dimensional spatial structure of Chinese characters and disassembles them into several parts divided by the radicals. Treating characters as a combination of multiple radicals instead of a single character reduces the number of characters that must be recognized. Through the combination of different radicals to identify characters that have never been encountered, Wu et al. [22] proposed the joint spatial and radical analysis network (JSRAN), whose architecture can effectively resolve the problems of huge characters and limited data when traditional Chinese character recognition is executed.

2.4. Transfer Learning

Transfer learning uses a pre-trained model, and then adopts a small amount of new data as input and retrain the model. Consequently, the pre-trained model learned to digest new information and infer what the old data do not. After several training cycles, the feature extraction method can effectively learn the characteristics of the new data in a short time, while retaining the feature extracted by the old data and identifying the new data. Qiao et al. [23] proposed a new Siamese network architecture. Traditionally, to learn new characters from existing models, they must use sufficient characters of data for training; however, they use a small amount of data for pre-training and learn new category characteristics by predicting the parameters of the activation function efficiently. Ao et al. [24] used hybrid models by combining RNN and CNN, when recognizing new handwritten characters, the characteristics of the word can be learned through the printed character and do not need the handwritten samples. Using printed characters as a prototype of handwritten characters and training the entire network, a small number of handwritten images are used for fine-tuning. Experimental results show that the proposed method can effectively use printed characters to learn the characteristics of handwritten characters. Tang et al. [25] utilized a large number of printed Chinese characters for model pre-training. After the training, its parameters were initialized as the parameters of the new model

158 and fine-tuned with a small amount of labeled ancient texts and handwritten Chinese
159 characters. The feature extractors and classifiers originally used in printed characters were
160 transferred and adapted to the new data. Tang et al. [26] proposed a semi-supervised
161 transfer learning (STL), which integrates the multi-kernel maximum mean discrepancy
162 (MK-MMD) loss function into the traditional transfer learning model, thereby narrowing
163 the gap between the source and target domains. They first used a large amount of labeled
164 data for CNN training, and then used a small amount of target domain data for fine-
165 tuning. Finally, a large number of unlabeled target domain data (with MK-MMD loss)
166 and a limited target domain of labeled data (with softmax loss) are used to train the entire
167 model simultaneously. It has been tested on several well-known CNN networks, including
168 AlexNet [27], GoogLeNet [5], and ResNet [28]. The experimental results show that the
169 proposed method effectively enhances the accuracy rate of Chinese character recognition.

170 **3. Methodology**

171 The main objective of this study was the construction of an automatic OCR system
172 designed to identify up to 13,070 large-scale printed Chinese characters using deep learn-
173 ing neural networks and fine-tuning techniques. The overall procedure involved in the
174 proposed framework is shown in Figure 2. The main steps are as follows. (1) Generate
175 a synthesized training set of character images with different Chinese character fonts and
176 simulated ID card backgrounds. (2) Perform data augmentation on the synthesized training
177 dataset by applying rotation and zooming to increase diversity. (3) Input the synthesized
178 training dataset to the proposed GoogLeNet-GAP model [5] for training; Finally, (4) collect
179 several samples of Chinese characters on real ID cards, perform data augmentation and
180 balancing processing for further transfer learning, and finally output the recognition results.
181 The steps executed are presented in the following subsections.

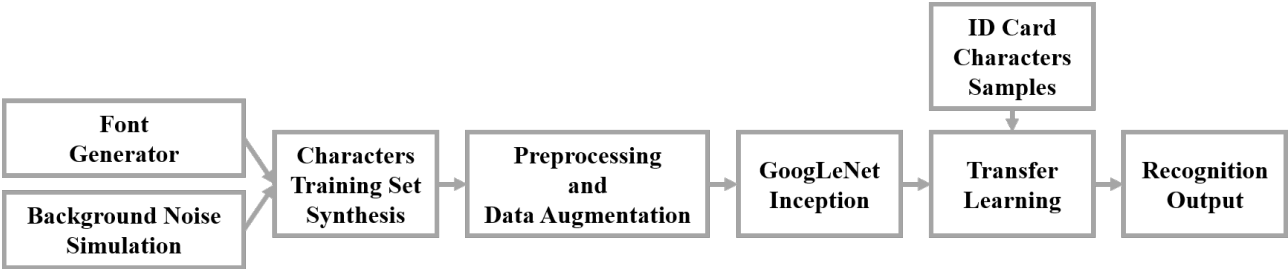


Figure 2. Flowchart of the proposed large-scale Chinese character recognition system.

182 **3.1. Training Dataset Synthesis**

183 The deep learning mechanism relies on a large amount of training data. However,
184 owing to the privacy issues associated with collecting personal ID card images, the avail-
185 ability of real data remains limited. Hence, the construction of large-scale synthetic data for
186 neural network training was necessary in this study. Furthermore, the appearance of text
187 on ID cards may vary owing to different backgrounds and lighting. Therefore, we created
188 the synthetic Chinese characters training dataset through the following steps. First, (1)
189 generate a font image with a white background. Then, (2) produce a simulated background,
190 and (3) combine the font and background images. The detailed data synthesis process is
191 presented below.

192 **Font Generation.** First, we used the Pillow [29] package in the Python programming
193 language to create a 100 × 100-pixel white image as the background and then pasted the
194 character text on the background to generate a character sample image. Big5 is a common
195 Chinese character encoding method used for traditional Chinese characters, which contains
196 a large set of 13,060 characters used in daily life. This study uses it as the classification
197 target and adds the digits 0-9 that often appear on the ID cards. There were a total of
198 13,070 characters. Because the most commonly used Chinese fonts are Microsoft JhengHei

and MingLiU, we chose these two fonts and added their boldface versions to increase the diversity of the dataset, with a total of four different fonts for each character. In addition, the Times News Roman font was adopted for the digits 0 to 9. Therefore, a total of 52,290 character images were generated.

Background Simulation. Second, to simulate the factors of the background stripes or anti-counterfeiting patterns on ID cards, we added noise and merged them into the aforementioned generated character images. In this study, we introduce three background simulation methods, including a random gray-level background, a random noise background, and a patch stitching background, as described below. The three proposed background simulation methods can effectively convey the background pattern of real ID cards and facilitate the learning capability of the CNN model to achieve better OCR performance.

- (1) To simulate various environmental lighting changes, we overlaid the generated character images with a different gray-level background, as depicted in Figure 3. The gray level was randomly selected. Additionally, the resultant character image was blurred with a Gaussian filter to smooth the appearance of the generated character image.



Figure 3. Background simulation method 1: Character image with the gray-level background.

- (2) The second type of background was random noise. First, a pure gray-level background was generated. Next, a positive or negative random number was added to the gray-level background. Then, the character image was merged with the noisy background, as shown in Figure 4. Finally, Gaussian blur was applied to smooth the appearance.

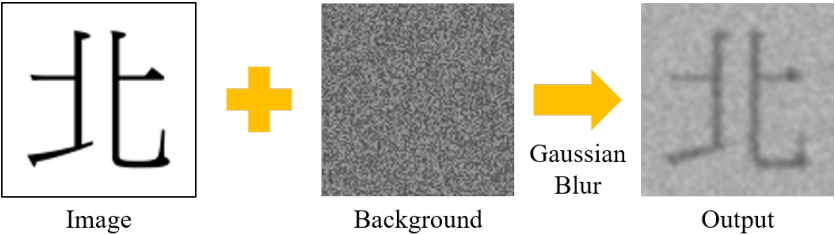


Figure 4. Background simulation method 2: Character image with random noise background.

- (3) To improve the instability caused by random noise, we further propose a patch stitching method to simulate the ID card background. First, we randomly selected 50 patches of 2×2 images from the surrounding areas of a real ID card image, as shown in the red boxes in Figure 5a. Next, these patches were stitched in order from left to right and from top to bottom into a 100×100 -pixel background image, as shown in Figure 5b. After obtaining the stitching background, it was combined with the character image, and Gaussian blurring was performed. Figure 6 illustrates the process.

3.2. Image Preprocessing and Data Augmentation

Grayscale Normalization. The problem of image brightness range variation is often encountered in the OCR task. To enhance the images, preprocessing of the brightness normalization techniques was applied. Based on the preliminary experiments using brightness

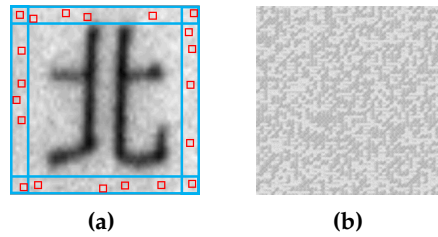


Figure 5. Stitching background method: (a) Random selection of patches, (b) Stitching background result.

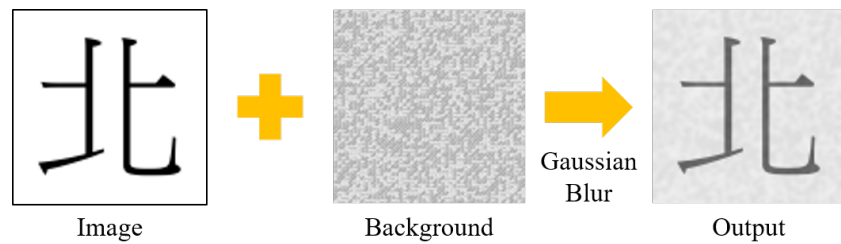


Figure 6. Background simulation method 3: Character image with patch-stitching background.

normalization techniques such as min-max, averaging, and histogram equalization, we adopted the min-max normalization method (Equation (1)) and achieved the best recognition accuracy. Figure 7 shows an example of min-max normalization. Compared to the original image (Figure (7a)), the normalized image appears to exhibit a better contrast. This enhancement leverages feature extraction for deep learning.

$$X_{minimax} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

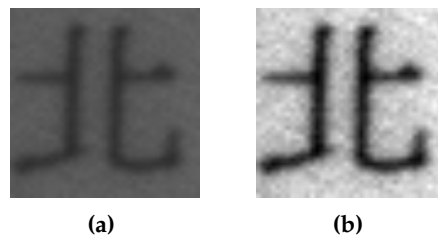


Figure 7. Example of normalization preprocessing: (a) Original image, (b) MiniMax normalization result.

Data Augmentation. To ensure that the proposed model can adapt to a variety of different situations, considering that the ID cards may be aligned in different angles or the ID card images may be captured at slightly different distances from the cameras, the rotation angle and size of the text image are expected to change accordingly. Data augmentation was further applied to the synthesized data generated in Section 3.1, where each image was randomly rotated between 10° and -10° , and enlarged between the ratios of 1.1 to 1.3, so that the diversity of the data increased. Examples of data augmentation images are shown in Figure 8.

3.3. Baseline Model Pre-Training

Using the different font generation, background simulation, and data augmentation processes illustrated in Sections 3.1 and 3.2, we expanded the dataset to generate a sufficient amount of training data, ultimately using up to more than 19.6 million images to accommodate the diversity of input image variations and strengthen the recognition

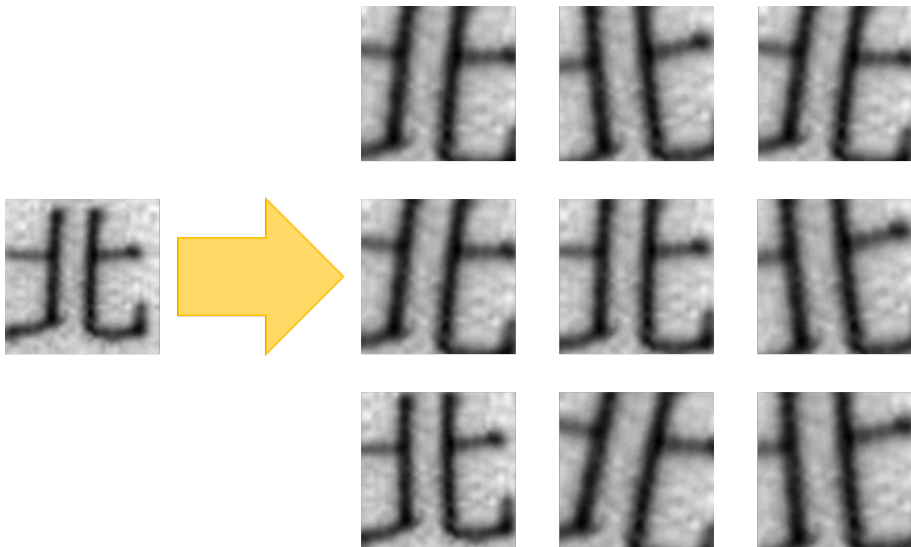


Figure 8. Examples of data augmentation images.

250 capability of the CNN model. In this study, we utilized and modified the GoogLeNet [5]
251 and MelnykNet [12] models as a baseline and pre-trained them with the abovementioned
252 synthetic Chinese characters dataset.
253 **GoogLeNet-GAP model.** GoogLeNet [5] was proposed by Google in 2014 and won
254 the ILSVRC competition. By adding different sizes of the kernel to extract the features
255 of images at different scales, they also added 1×1 convolution to reduce the amount
256 of calculation, and finally combined these features to form the Inception v1 block [5].
257 Combining multiple inception blocks can increase the diversity of the feature map and
258 make the model converge more quickly. Because our goal is to classify large-scale categories
259 of Chinese printed characters, to avoid overfitting caused by a large number of characters,
260 we replaced the FC layer with a global average pooling (GAP) layer (as plotted in red dash
261 line in Figure 9) to reduce the number of model parameters and improve its classification
262 performance. Figure 9 shows the modified network architecture GoogLeNet-GAP.

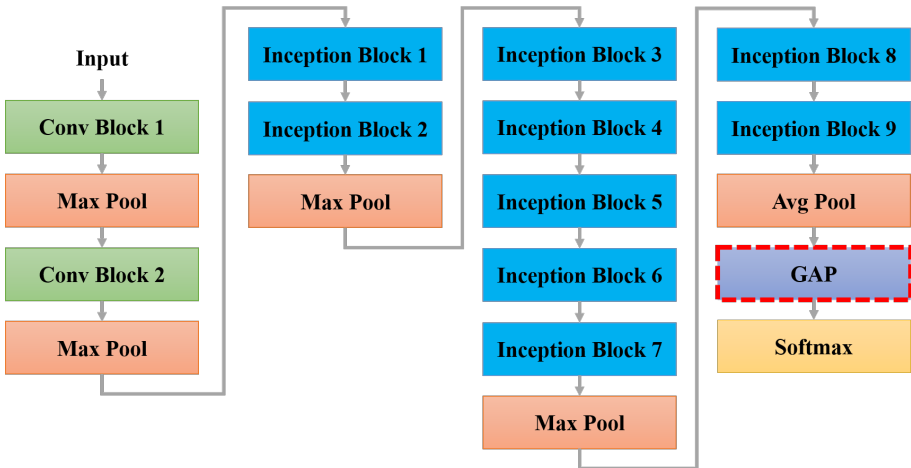


Figure 9. Modified GoogLeNet architecture: GoogLeNet-GAP model.

263 **MelnykNet-Res model.** MelnykNet was originally proposed for offline handwritten
264 Chinese character recognition [12]. In the study of Melnyk et al., the handwritten character
265 data were binary images and the background was clear [12]; hence, the MelnykNet model
266 only requires a few parameters and can be trained to perform well. However, as our input
267 images are grayscale images with a noisy background, it was necessary to strengthen the
268 feature extraction capabilities of the CNN model to achieve better recognition performance.

269 According to Matthew et al. [30], the feature maps extracted by each convolutional layer
270 differ; the higher the layer, the finer the feature maps, and vice versa. In the original
271 architecture, the contour of the characters were extracted by the previous convolutional
272 layer, so we added a residual block [28] (as plotted in red dash line in Figure 10) to the later
273 layer to continue to extract the detailed features of the character. The modified MelnykNet
274 model MelnykNet-Res is illustrated in Figure 10.



Figure 10. Modified MelnykNet architecture: MelnykNet-Res model.

275 In this study, we collected a test dataset containing 4,944 samples of 294 Chinese char-
276 acters cropped from real ID card images. Preliminary experiments were conducted using
277 the abovementioned GoogLeNet-GAP model and MelnykNet-Res model. We performed
278 training using the synthetic training dataset and compared the recognition rates of the
279 two models using a real testing dataset. Table 2 presents the performance of these two
280 models, given the different numbers of training samples augmented for each character.
281 As may be observed from Table 2, the modified GoogLeNet model performed better than
282 the MelnykNet model. The highest recognition rate was 95.71%. Therefore, we chose the
283 modified GoogLeNet-GAP model for further study.

Table 2. Performance comparison of the modified MelnykNet-Res model and GoogleLeNet-GAP model tested on real ID card data.

# of Training Samples per Character	MelnykNet-Res Accuracy	GoogLeNet-GAP Accuracy
300	21.64%	51.40%
600	41.75%	90.62%
900	53.40%	95.71%
1500	68.95%	95.35%

284 3.4. Mixed Data Model and Transfer Learning Model

285 Although the proposed GoogLeNet-GAP model achieved good recognition accuracy
286 on the real ID card character testing dataset (see Section 3.3), a certain degree of mis-
287 classification remained. To improve the performance of our model, we expanded the
288 synthetic training set by including an augmented version of the data from a small part of
289 the real ID card character images. We propose two new models: (1) a mixed data model
290 and (2) a transfer learning model. The mixed data model was obtained by retraining the
291 GoogLeNet-GAP model (see Figure 9) from scratch while mixing synthetic training data
292 and augmentation data of partial samples from real ID card character images. In contrast,
293 the transfer learning model used the pre-trained model (see Section 3.3) as the base model
294 and performed fine-tune training with the augmented data of partial samples from the
295 real ID card character images. The abovementioned two methods were validated through
296 experiments in which both the proposed new models achieve substantial improvements in
297 recognition accuracy. Moreover, the results show that the transfer learning model is more
298 practical for practical applications. The model recognition performance can be gradually
299 improved by periodically fine-tuning with newly collected real data. Meanwhile, the
300 training speed can dramatically increase because of the very small amount of new data.
301 Section 4 presents the detailed experimental results and reveals the advantages of the
302 transfer learning model.

303 **4. Experimental Results**

304 *4.1. Dataset and Implementation Details*

305 As illustrated in Section 3.1, the training dataset was constructed by synthesizing five
306 different fonts of the common Chinese character Big5 code and digits using the Pillow [29]
307 package in Python. A total of 52,290 character images were generated. Furthermore, to
308 accommodate the diversity of input image variations, each of these 52,290 characters was
309 synthesized again with different noisy backgrounds or augmented with the angle rotation
310 and resizing processes described in Sections 3.2. Finally, we expanded the dataset to
311 generate a sufficient amount of training data, ultimately including up to more than 19.6
312 million images. Table 3 lists the detailed content of the dataset mentioned in this study. To
313 the best of our knowledge, no dataset with a sufficient number of Chinese character images
314 to be used for a large-scale classifier has been developed in prior work. The proposed
315 synthesis method can automatically generate character images through programs and
316 can thus create a sufficient amount of training data and include augmentation to increase
317 diversity.

Table 3. Comparison of the common Chinese character datasets and the proposed dataset.

Dataset	# of Characters	# of Images	Type
CASIA-HWDB1.0	3,866	1,609,136	Simplified Chinese
CASIA-HWDB1.1	3,755	1,121,749	Simplified Chinese
CASIA-HWDB1.2	3,319	990,989	Simplified Chinese
Chinese MNIST	13,065	587,925	Traditional Chinese
CASIA-AHCDB	10,350	more than 2.2 million	Traditional Chinese
Ours	10,370	19.6 million	Traditional Chinese

318 The synthetic training dataset was divided into three parts; 70%, 10%, and 20% were
319 used for network training, validation, and testing, respectively. The validation set was used
320 to avoid overfitting during training, while the testing set was used to verify the effects of the
321 trained model. Our models were implemented in Python 3.7 on the TensorFlow platform
322 under a Windows operating system. The experiments were conducted on workstation
323 computer with an Intel Core i9-9900X CPU at a clock rate of 3.50GHz with 64GB of RAM
324 and an NVIDIA GTX 2080 Ti GPU. All the character recognition performances presented in
325 this study were evaluated with the same testing dataset collected from real ID card images,
326 which contained 4,944 samples of 294 Chinese characters. The image size of each sample is
327 around 45×45 cropped from the ID card images. The ID card images were captured using
328 the device shown in Figure 1.

329 *4.2. Ablation Study*

330 To verify that our GoogLeNet-GAP network can overcome the overfitting problem
331 and obtain higher recognition accuracy, we compared the performances of the models with
332 the original FC layer and the new GAP layer by providing a different number of training
333 samples per character for the 294 output classes in the GoogLeNet model. Table 4 shows
334 the effect of using the GAP layer as the classification output layer. Our GoogLeNet-GAP
335 network achieved higher accuracy. Moreover, it may be observed that when provided
336 additional training samples, the network model performed better. Therefore, subsequent
337 experiments will be based on the GoogLeNet-GAP model.

338 In addition, for the fully connected nature of FC, when the number of character
339 categories increased, the number of FC layer parameters also increased dramatically. Thus,
340 the chances of overfitting increased sharply. When using GAP, because its hidden layer
341 does not have trainable parameters, the parameters of the output layer are only increased
342 on increasing the number of classifications. Therefore, compared with FC, GAP has fewer
343 parameters and is less likely to cause overfitting. Furthermore, the use of fewer parameters
344 can increase the computational speed of the CNN. Table 5 shows the number of parameters
345 used in the FC and GAP for various classification tasks.

Table 4. Performance comparison of GoogLeNet models using FC and GAP.

# of Training Samples per Character	FC	GAP
300	41.59%	51.40%
600	87.74%	90.62%
900	85.56%	95.71%
1500	91.20%	95.35%

Table 5. Comparison of parameters size of FC and GAP in GoogLeNet model.

# of Classes (Characters)	# of FC Parameters (million)	# of GAP Parameters (million)
294	11.802	11.716
1812	16.557	13.271
4945	40.941	16.483
13070	195.649	24.811

According to the Ministry of Education of Taiwan, 4,803 standard Chinese characters are frequently used in daily life. There are 1,802 characters commonly used in addresses according to the postal system. In addition to our target of the large-scale classification of 13,070 characters, we also conducted experiments on the models of 1,812 outputs (1,802 characters plus 10 digits) and 4,945 outputs (the union of 4,803 standard Chinese characters, 1,802 characters in postal system, and 10 digits) for the postal system and education system, respectively. Table 6 indicates that the GoogLeNet-GAP maintained classification performance, even when the number of model outputs was expanded from 1,812 to 13,070. Furthermore, all the character recognition tasks were performed with an accuracy exceeding 90% when the number of training samples per character increased to 900. The system achieved the best performance on average, given 1,500 training samples per character. Nevertheless, the more training samples provided, the more time required to train the model. If the model can be further improved to achieve compatible performance with less training data, it will become more practical for real-time applications.

Table 6. Performance of various large-scale GoogLeNet-GAP models with different numbers of training samples per character.

# of Training Samples per Character	294 Classes (Characters)	1,812 Classes (Characters)	4,945 Classes (Characters)	13,070 Classes (Characters)
300	51.40%	77.91%	76.52%	66.51%
600	90.62%	88.21%	89.99%	83.72%
900	95.71%	91.91%	89.42%	88.51%
1,500	95.35%	95.28%	90.23%	92.17%

4.3. Mixed Data Model and Transfer Learning Model Results

4.3.1. Mixed Data Model

The mixed data model was obtained by retraining the GoogLeNet-GAP model from scratch. The training set was composed of the augmented versions of partial real data and partial synthetic training data created previously (see Sections 3.1 and 3.2). We randomly selected 66 of the 294 characters from the real data. Then, rotation and resizing augmentation are applied to the original 235 images of these 66 characters and then expanded 20% of the number of training samples of each character. For instance, if the number of training samples per character was 600, then 120 samples were obtained from real data augmentation and 480 samples were taken from the synthetic dataset of the corresponding character. Finally, the model was re-trained. Table 7 shows that the accuracy increased

371 significantly, not only for the effect of the original 66 characters but also for the new char-
372 acters. Although the results obtained by this method were good, there were still certain
373 drawbacks. Once new real data are available, the model must be re-trained after the new
374 data are mixed with the original data. This causes a substantial increase in training time,
375 and thus, this function cannot be used for real-time recognition services.

Table 7. Performance comparison of the original model and the mixed data model with 600 and 900 training samples per character.

# of Classes (Characters)	600 Training Samples per Character		900 Training Samples per Character	
	Original Model	Mixed Data Model	Original Model	Mixed Data Model
294	90.62%	96.84%	95.71%	98.51%
1812	88.21%	96.22%	91.91%	95.03%
4945	89.99%	93.54%	89.42%	96.16%
13070	83.72%	91.95%	88.51%	93.93%

376 4.3.2. Transfer Learning Model

377 One of the main advantages of transfer learning is that a pre-trained model can
378 quickly adapt to new data through fine-tuning training using small amounts of real data.
379 To evaluate the performance of transfer learning, we fine-tuned various classification
380 models with different amounts of real data by providing 10 and 20 augmentation samples
381 per character. However, if only a small number of characters are used to fine-tune a large-
382 scale character classification, the model becomes biased for a small number of characters
383 and thus loses the recognition rate for other characters. To avoid deviations, data balancing
384 is applied by recruiting the synthetic data generated in Section 3.1 for the remaining
385 characters and then performing training. Table 8 presents the results of fine-tuning using
386 three different numbers of real character data (100, 200, and 294) for two training scenarios,
387 namely, 10 and 20 samples per character. As can be observed from Table 8, our transfer
388 model can improve the recognition accuracy of the characters on printed ID cards with
389 fine-tuning. After the data are generated and balanced, the problems of insufficient data
390 and learning deviation are solved simultaneously. Figure 11 shows that the character
391 recognition performance was improved dramatically when the models were fine-tuned
392 using more real data.

Table 8. Testing results on transfer model fine-tuned using three different numbers of real character data (100, 200, and 294) for two training scenarios, namely, 10 and 20 projection real samples per character.

# of Real Samples per Characters	# of Output Classes	# of Characters Fine-tuned in Transfer Learning			
		0	100	200	294
10	294	90.62%	98.95%	99.33%	99.35%
	1812	88.21%	98.63%	98.87%	99.35%
	4,945	89.99%	98.24%	98.63%	99.23%
	13,070	83.72%	97.74%	98.67%	98.85%
20	294	90.62%	98.75%	99.45%	99.33%
	1812	88.21%	98.58%	98.83%	99.41%
	4,945	89.99%	98.87%	98.93%	99.47%
	13,070	83.72%	97.34%	98.85%	99.39%

393 4.4. Error Analysis and Performance Enhancement

394 In Section 4.3, we demonstrate that the two proposed transfer learning methods
395 were able to strengthen the model trained using synthesized data and real data. The
396 experimental results show that the recognition rate was improved, and it reached more
397 than 95% after adding a few real samples for transfer learning. Although our model

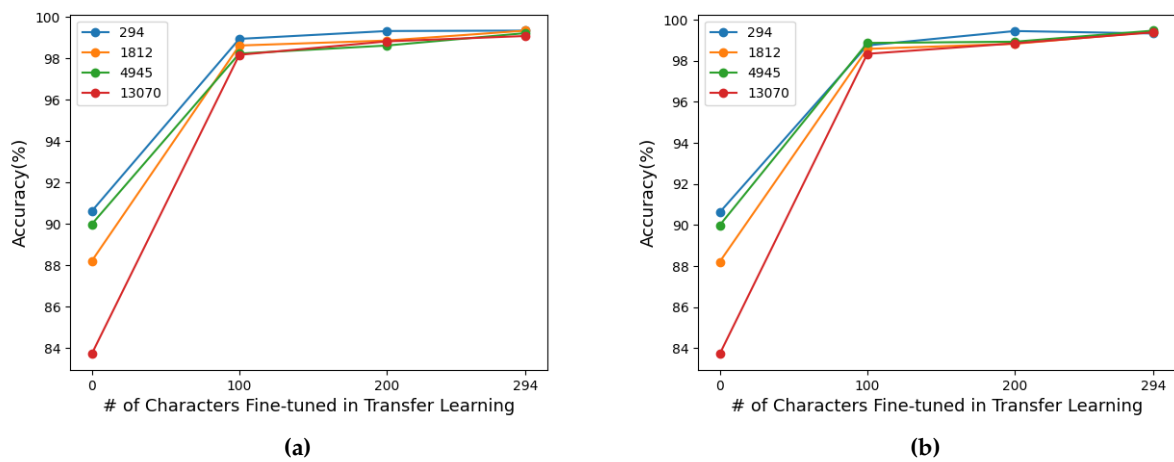


Figure 11. Character recognition performance was improved dramatically when the models were fine-tuned using three different numbers of real character data (100, 200, and 294) for two training scenarios, namely, (a) 10 real samples per character, and (b) 20 real samples per character.

398 achieved good recognition results under large-scale classifications, some issues remain
 399 nonetheless. Figure 12 shows some samples with incorrect recognitions in the large-scale
 400 classification of 13,070 characters. There are some Chinese characters that seem to be similar;
 401 however, in reality, they have completely different meanings. On ID cards with a cluttered
 402 background, the text is more likely to be disturbed by background noise. In addition, an
 403 ID card may be stained or damaged due to wear, which may result in misclassifications
 404 by the CNN model. Figure 12a displays two examples of misclassification, as mentioned
 405 above. In addition, if the text segmentation method is not sufficiently robust, it may easily
 406 cause the cropped text image to contain more than one character, or to retain only parts of
 407 the character. Figure 12b presents two examples of misclassification due to inappropriate
 408 cropping.

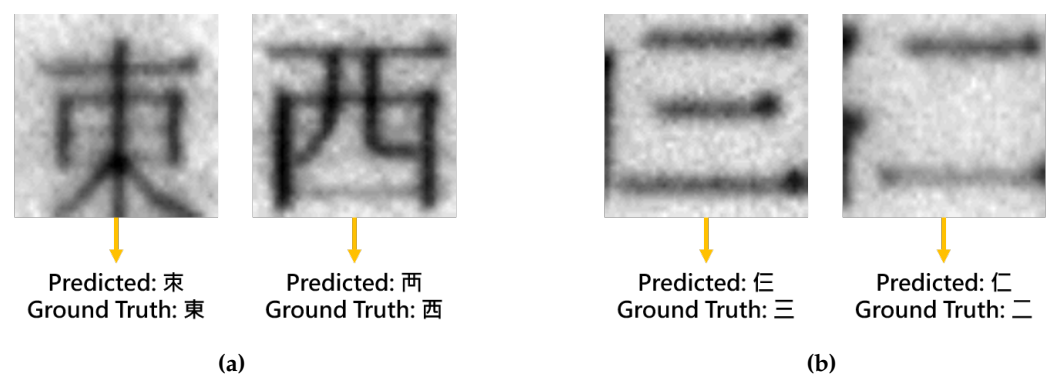


Figure 12. Some error misclassification samples: (a) Contaminated and similar error, (b) Cropping error.

409 To solve the recognition problem caused by character segmentation shifting, we
 410 utilized the projection method. First, the grayscale character image was converted to a
 411 binary image with a specific threshold to separate the character from the background.
 412 Then, binary pixels were projected in the horizontal and vertical directions separately to
 413 locate the area of the character in the image. However, the ID card images may include a
 414 noisy background, and the projection may be affected by noise disturbances. Therefore, we
 415 performed morphology processing using a 1×7 vertical bar-type erosion operator on the
 416 horizontal projection image and disconnected the small adjacent blocks. Next, we search

to check if there was a continuous black area from the boundary toward the center, as indicated by the red lines in Figure 13. Meanwhile, the distance between the two red lines was required to be greater than 60% of the original width or height to enclose a complete character. As shown in Figure 13, a better segmentation result was obtained. Next, the OCR system was re-evaluated using the rectified dataset.

Table 9 presents the new evaluation results of the proposed models with the 13,070 character classification task. It may be observed that the recognition accuracy was significantly improved after projection rectification. The model performance was further improved by incorporating the rectified real dataset into the learning process for both the mixed data model and the transfer learning model.

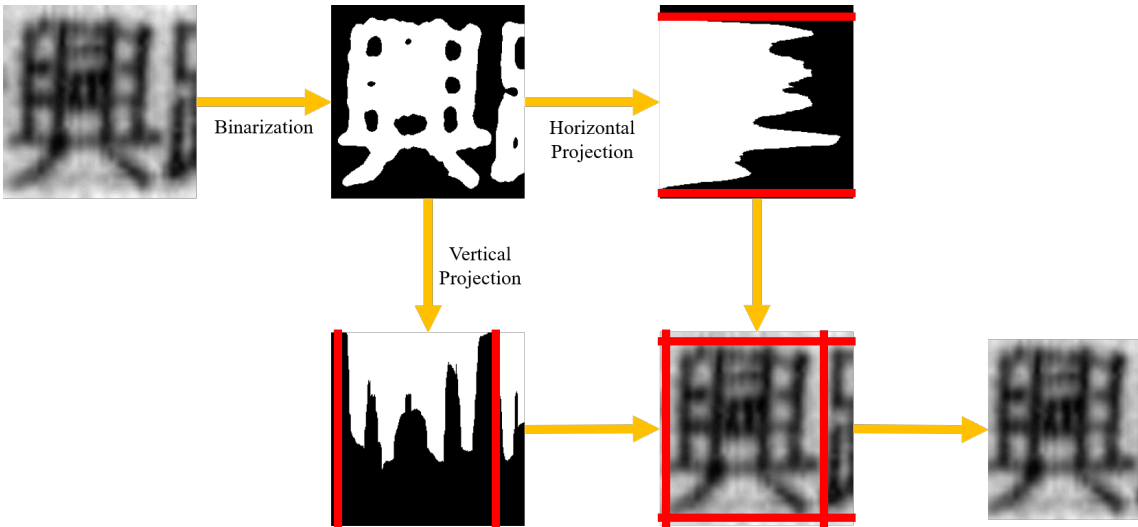


Figure 13. Procedure of projection method.

Table 9. New evaluation results of the proposed models on the 13,070 characters classification task using the projection rectification testing set.

Model	Training Samples	Original	Projection
Font model	600	83.72%	85.70%
Mixed model	600	91.95%	96.54%
Transfer model	400	95.01%	97.53%

5. Conclusion

The remarkable and rapid development of deep learning technology has achieved great successes in computer vision and has also played a pivotal role in related fields. In this study, we have developed an automatic OCR system to identifying up to 13,070 large-scale printed Chinese characters by using deep learning neural networks and fine-tuning techniques. The proposed framework comprises four components: training dataset synthesis, background simulation, image preprocessing and data augmentation, model training, and transfer learning. Specifically, the training data synthesis procedure was composed of character font generation and a background simulation process. Three background models were proposed to simulate the background noise and anti-counterfeiting patterns on ID cards. Subsequently, the character font text is pasted on various backgrounds to generate character sample images. The preprocessing and data augmentation module first performs the min-max normalization operation to consistently rescale the brightness of the character images. Then, rotation and zooming data augmentation were applied to the synthesized training dataset to expand the diversity. A massive dataset of more than 19.6 million images

was created to accommodate the variations of input images and strengthen the learning capacity of the CNN model. The proposed data generation method has been validated experimentally to simulate the text data on ID cards and use a consistent normalization process to improve the brightness and contrast of the original image, making the model more adaptable to different backgrounds on ID card characters. In the deep learning network design, we modified GoogLeNet by replacing the FC layer with a GAP layer to avoid overfitting caused by a large amount of training data. The number of model parameters was consequently reduced. Finally, we employed the transfer learning technique to further refine the CNN model using a very small number of real data samples. The two transfer learning models we proposed can improve the learning of the original model within an acceptable range. Through the usage of real data, the proposed approach can be adapted to the characters on ID cards. Furthermore, the input character images were further rectified by applying the projection method. After the implementation of the data balance, transfer learning can be performed from each category on average, instead of only targeting a few characters, which considerably reduces the instability caused by the use of transfer learning under a large-scale classification. Overall, the overall recognition performance improved significantly. The experimental results demonstrate that the proposed framework is effective, and the accuracy of the large-scale 13,070 character recognition system was as high as 99.39% when evaluated on our real ID card dataset.

Although our model exhibited good results on the identification of characters on ID cards, certain characters were nonetheless recognized incorrectly; in particular, there are quite a few Chinese characters that seem to be similar, but in fact they are not the same, and their meanings are quite different. In addition, there is still a certain degree of difference between the characters on the ID cards and the simulated data generated by us. In the future, we hope to collect more samples of Chinese characters from ID cards for fine-tuning training. We are currently deploying the proposed OCR framework to a real-time identification device aimed at reducing the manpower demand of the service industry. Accurate automated identification can also reduce the possibility of human error. In addition, we intend to compare the accuracy of the current state-of-the-art methods with our proposed method under various environments and operations; we also aim to improve the objective validation of our method and the OCR performance. Lastly, it would be beneficial to explore additional language systems and design the system to recognize all types of ID card text.

Author Contributions: Conceptualization, Yi-Quan Li, Hao-Sen Chang and Daw-Tung Lin; Data curation, Yi-Quan Li and Hao-Sen Chang; Formal analysis, Yi-Quan Li and Daw-Tung Lin; Funding acquisition, Daw-Tung Lin; Investigation, Yi-Quan Li and Daw-Tung Lin; Methodology, Yi-Quan Li, Hao-Sen Chang and Daw-Tung Lin; Project administration, Yi-Quan Li; Resources, Yi-Quan Li and Hao-Sen Chang; Software, Hao-Sen Chang; Validation, Yi-Quan Li and Hao-Sen Chang; Writing – original draft, Hao-Sen Chang; Writing – review and editing, Daw-Tung Lin. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Ministry of Science and Technology, Taiwan, Grant MOST 110-2622-E-305-001 and by the Orbit Technology Incorporation..

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

Data Availability Statement: The experiment dataset reported in this paper is available at <http://imslab.csie.ntpu.edu.tw/index.php/dataset>.

References

1. Aprillian, H.D.D.; Purnomo, H.D.; Purwanto, H. Utilization of Optical Character Recognition Technology in Reading Identity Cards. *International Journal of Information Technology and Business* **2019**, *2*, 38–46.
2. Purba, A.M.; Harjoko, A.; Wibowo, M.E. Text Detection In Indonesian Identity Card Based On Maximally Stable Extremal Regions. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)* **2019**, *13*, 177–188.

3. Satyawan, W.; Pratama, M.O.; Jannati, R.; Muhammad, G.; Fajar, B.; Hamzah, H.; Fikri, R.; Kristian, K. Citizen Id Card Detection using Image Processing and Optical Character Recognition. *Journal of Physics: Conference Series*. IOP Publishing, 2019, Vol. 1235, p. 012049.
4. Tavakolian, N.; Nazemi, A.; Fitzpatrick, D. Real-time information retrieval from Identity cards. *arXiv preprint arXiv:2003.12103* **2020**.
5. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.
6. Xu, Y.; Shan, S.; Qiu, Z.; Jia, Z.; Shen, Z.; Wang, Y.; Shi, M.; Eric, I.; Chang, C. End-to-end subtitle detection and recognition for videos in East Asian languages via CNN ensemble. *Signal Processing: Image Communication* **2018**, *60*, 131–143.
7. Zhong, Z.; Jin, L.; Xie, Z. High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 846–850. doi:10.1109/ICDAR.2015.7333881.
8. Yin, F.; Wang, Q.F.; Zhang, X.Y.; Liu, C.L. ICDAR 2013 Chinese Handwriting Recognition Competition. 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 1464–1470. doi:10.1109/ICDAR.2013.218.
9. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv preprint arXiv:1312.4400* **2013**.
10. Li, Z.; Teng, N.; Jin, M.; Lu, H. Building efficient CNN architecture for offline handwritten Chinese character recognition. *International Journal on Document Analysis and Recognition (IJ DAR)* **2018**, *21*, 233–240.
11. Xiao, X.; Jin, L.; Yang, Y.; Yang, W.; Sun, J.; Chang, T. Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition. *Pattern Recognition* **2017**, *72*, 72–81.
12. Melnyk, P.; You, Z.; Li, K. A high-performance CNN method for offline handwritten Chinese character recognition and visualization. *Soft Computing* **2019**, *24*, 7977–7987. doi:10.1007/s00500-019-04083-3.
13. Su, Y.S.; Chou, C.H.; Chu, Y.L.; Yang, Z.Y. A Finger-Worn Device for Exploring Chinese Printed Text With Using CNN Algorithm on a Micro IoT Processor. *IEEE Access* **2019**, *7*, 116529–116541. doi:10.1109/ACCESS.2019.2936143.
14. Liu, B.; Xu, X.; Zhang, Y. Offline Handwritten Chinese Text Recognition with Convolutional Neural Networks. *arXiv preprint arXiv:2006.15619* **2020**.
15. Chen, P.C. Traditional Chinese Handwriting Dataset. <https://github.com/AI-FREE-Team/Traditional-Chinese-Handwriting-Dataset>, 2020.
16. Xu, Y.; Yin, F.; Wang, D.H.; Zhang, X.Y.; Zhang, Z.; Liu, C.L. CASIA-AHCDB: A Large-Scale Chinese Ancient Handwritten Characters Database. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 793–798. doi:10.1109/ICDAR.2019.00132.
17. Zhong, Z.; Jin, L.; Feng, Z. Multi-font printed Chinese character recognition using multi-pooling convolutional neural network. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 96–100. doi:10.1109/ICDAR.2015.7333733.
18. Qiu, S. Global weighted average pooling bridges pixel-level localization and image-level classification. *arXiv preprint arXiv:1809.08264* **2018**.
19. Zhang, Q.; Lee, K.C.; Bao, H.; You, Y.; Li, W.; Guo, D. Large Scale Classification in Deep Neural Network with Label Mapping. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 1134–1143. doi:10.1109/ICDMW.2018.00163.
20. Zhang, J.; Zhu, Y.; Du, J.; Dai, L. Radical Analysis Network for Zero-Shot Learning in Printed Chinese Character Recognition. 2018 IEEE International Conference on Multimedia and Expo (ICME), 2018, pp. 1–6. doi:10.1109/ICME.2018.8486456.
21. Zhang, J.; Du, J.; Dai, L. Radical analysis network for learning hierarchies of Chinese characters. *Pattern Recognition* **2020**, *103*, 107305.
22. Wu, C.; Wang, Z.R.; Du, J.; Zhang, J.; Wang, J. Joint Spatial and Radical Analysis Network For Distorted Chinese Character Recognition. 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019, Vol. 5, pp. 122–127. doi:10.1109/ICDARW.2019.40092.
23. Qiao, S.; Liu, C.; Shen, W.; Yuille, A. Few-Shot Image Recognition by Predicting Parameters from Activations. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7229–7238. doi:10.1109/CVPR.2018.00755.
24. Ao, X.; Zhang, X.Y.; Yang, H.M.; Yin, F.; Liu, C.L. Cross-Modal Prototype Learning for Zero-Shot Handwriting Recognition. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 589–594. doi:10.1109/ICDAR.2019.00100.
25. Tang, Y.; Peng, L.; Xu, Q.; Wang, Y.; Furuhashi, A. CNN Based Transfer Learning for Historical Chinese Character Recognition. 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 2016, pp. 25–29. doi:10.1109/DAS.2016.52.
26. Tang, Y.; Wu, B.; Peng, L.; Liu, C. Semi-Supervised Transfer Learning for Convolutional Neural Network Based Chinese Character Recognition. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, Vol. 01, pp. 441–447. doi:10.1109/ICDAR.2017.79.
27. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **2012**, *25*, 1097–1105.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
29. Clark, A. Pillow (PIL Fork) Documentation, 2015.

30. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. European Conference on Computer Vision (ECCV). Springer, 2014, pp. 818–833.