Vision Transformer-Based Recognition of Chinese Cursive Calligraphy: A Curriculum Learning and Skeleton Embedding Approach

Xinrui Shan[1], Jinyang Zheng[2], Yilin Fang[3], Tianhong Qi[4]
{xinruishan@zju.edu.cn[1], Jinyang.Zheng22@xjtlu.edu.cn[2],
fangyilin2023@bupt.edu.cn[3], U202241632@xs.ustb.edu.cn[4]}

College of Computer Science and Technology, Zhejiang University, Hangzhou, China[1]
School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China[2]
International School, Beijing University of Posts and Telecommunications, Beijing, China[3]
School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing, China[4]

Abstract. Chinese cursive calligraphy, characterized by fluid and complex strokes, presents a significant challenge in character recognition due to the variations in character structure and style. This paper proposes an innovative approach to recognize Chinese cursive characters using two Vision Transformer (ViT)-based models. We enhance the models with curriculum learning to optimize training efficiency by dynamically adjusting the difficulty of samples, allowing the models to progressively learn from easier to harder examples. Additionally, we integrate skeleton embeddings into the ViT encoder input to capture the underlying structural information of cursive characters. Our method demonstrates superior performance compared to baseline approaches, achieving higher recognition accuracy on a self-made cursive calligraphy datasets.

Keywords: Vision Transformer, Chinese Cursive Calligraphy, Skeleton Embedding

1 Introduction

The vast number and diversity of Chinese characters present significant challenges for recognition systems. The commonly used character set includes over 6000 characters, with an even larger extended character set [1]. This results in an exceptionally large number of categories. Additionally, the complex and varied structures of Chinese characters, characterized by a high average number of strokes. Many different characters exhibit similar spatial structures and stroke patterns, and same character has different handwriting styles, leading to confusion during recognition.

Despite years of research, there are still significant challenges in offline handwritten Chinese character recognition tasks, especially regarding unconstrained handwritten font recognition tasks [2].

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354609

In the past, handwritten Chinese character recognition methods have typically been categorized into three approaches: holistic methods, radical-based methods, and stroke-based methods [3]. These methods have achieved considerable success. For example, Tang et al. [4] proposed Convolutional Neural Networks (CNNs) that automatically extract high-level features from character images through multiple convolutional and pooling layers for classification. He et al. [5] introduced Deep Residual Networks (ResNet), which incorporate residual connections to address the degradation problem in deep networks, thereby improving recognition performance. However, due to the free and unrestrained calligraphy structure of cursive fonts, the development of Chinese cursive character detection, has been relatively slow [6].

As a treasure of Chinese calligraphy, cursive script is renowned for its unique writing style and elegant brushstrokes. However, its complex structures pose significant challenges to the automatic recognition of computers. In addition, there may be significant differences in alignment between cursive scripts using the same Chinese characters, which could significantly affect the accuracy of predictions. In recent years, advances in computer vision technology and deep learning algorithms have been able to meet the demand for automatic recognition of various complex fonts. In this context, the Visual Transformer (ViT) model has demonstrated outstanding performance in numerous visual tasks due to its powerful representation capabilities and parallel processing advantages. However, on the one hand, due to the complexity of cursive font structure, using ViT alone cannot achieve a desirable performance. On the other hand, in the current task of using CNN for skeleton-based Chinese character recognition. Tang et al. [4] found that extracting the skeleton structure of Chinese characters is an effective method for enhancing datasets. However, ViT did not consider enhancing feature extraction by this crucial step.

Therefore, we attempt to transfer this approach to the recognition of cursive script. In this paper, we propose a curriculum learning-based method as shown in the figures 1. Initially, the model focuses on processing easy samples, gradually transitioning to more complex tasks. By training the model in stages, we can incrementally enhance its capability to handle intricate samples. Moreover, to elevate recognition accuracy, our proposed method incorporates a novel approach of skeletonization applied to cursive script images. Specifically, for cursive script images in the dataset, we retained only the geometric and topological structures of the original characters while eliminating irrelevant information and noise. To further emphasize the structural features of the characters, we preserve patches from the original images and superimpose these patches onto corresponding hollowed-out cursive script image patches.

Ultimately, these superimposed patches are embedded into the embedding layer of the model, optimizing feature extraction and representation, and achieving feature fusion within the encoder embedding.

Thus, our research aims to integrate curriculum learning and skeleton embedding into the ViT model to better improve the accuracy of cursive script recognition tasks. By this method, we have made significant progress in the task of Chinese Cursive Calligraphy recognition, demonstrating the broad application potential of this method in complex character recognition. This innovative application in the field of handwritten Chinese character recognition not only helps enhance the recognition accuracy of cursive characters but also provides new methods for recognizing other complex fonts.

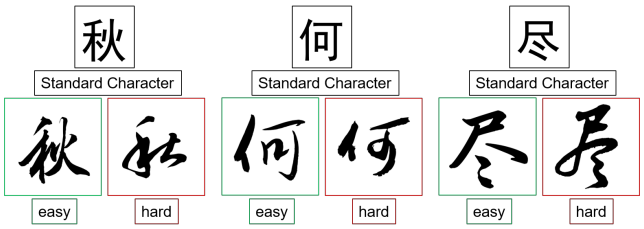

Fig. 1. Three sets of Chinese character images are displayed, a standard character, and its corresponding easy and hard samples.

The main contributions of our paper are summarized below:

- We apply Curriculum Learning to mimic the human learning process. Starting with the entire dataset, we first focus on simple samples, gradually introducing complex ones to enhance learning efficiency and performance.
- We introduce Skeleton Embedding, combining structural information of the image with original embeddings to boost the Vision Transformer's recognition capability.
- We constructed a dataset containing 10,180 images of single character cursive script. We enhanced ViT's performance through data augmentation and skeleton extraction, further optimizing training with Curriculum Learning under specific conditions.

2 Related Work

2.1 Traditional Vision Model in Character Recognition

In the domain of character recognition, traditional vision models have made significant contributions, particularly to the recognition of complex Chinese scripts. Liao et al. [7] enhanced the M6 model (EM6) by incorporating batch normalization and an additional fully connected layer, demonstrating the effectiveness of deep neural networks in recognizing cursive Chinese calligraphy. Additionally, Li et al. [8] addressed the challenges of the one-to-many relationship between stroke sequences and Chinese characters by proposing a zero-shot stroke-based method, which effectively recognizes handwritten, artistic, and scene characters.

The integration of classical image processing techniques with deep learning models has also shown promise. Si et al. [9] combining Histogram of Oriented Gradients (HOG) feature extraction, Euler distance calculation, and the Inception-v3 model to achieve high recognition accuracy for Chinese calligraphy characters. Similarly, Jin et al. [10] developed an offline handwritten Chinese character recognition method using Character-SIFT and an elastic grid partitioning technique.

Addressing the challenges of cursive script recognition, Qin et al. [11] created a dataset of cursive images and improved the SE-seglink method by incorporating Squeeze-and-Excitation operations to enhance feature extraction for cursive scripts. Furthermore, Zhong et al. [12] combined

Spatial Transformer Networks (STN) and Deep Residual Networks (DRN) to handle handwritten Chinese characters with varying positions, scales, and orientations. Our work integrates curriculum learning with skeleton embedding for feature extraction, with the aim of further improving the accuracy and adaptability of models to recognize cursive characters across diverse conditions.

Focusing on handwritten character recognition in different languages, Shanthi et al. [13] developed a system for Tamil characters using optimized preprocessing techniques and pixel density-based feature extraction methods combined with Support Vector Machine (SVM) classifiers. Shi et al. [14] introduced the CRNN model to efficiently process image sequences of varying lengths, particularly excelling in scene text recognition. Further enhancing robustness in text recognition, Shi et al. [15] developed RARE which uses spatial transformation networks (STN) to correct text images and attention-based sequence recognition networks (SRN) to accurately recognize irregular text. Additionally, Ptucha et al. [16] proposed a fully convolutional neural network architecture that directly outputs symbol streams from handwritten text images, eliminating the need for complex sequence alignment.

2.2 Transformer-based in Character Recognition

With the advancement of deep learning technologies, particularly the successful adoption of the Transformer architecture, the Vision Transformer (ViT) [17] model has become a pivotal research focus in image recognition. Its robust performance across various visual tasks has sparked significant interest among researchers in exploring its potential applications in character recognition.

The Transformer-based approaches have shown remarkable progress in this field. For instance, Xie et al. [18] achieved a breakthrough in artistic text recognition by integrating corner point features, character contrastive loss, and Transformer. Similarly, Gan et al. [19] proposed the PyGT method, which combines Transformer with Graph Convolutional Networks (GCN) to enhance the accuracy of handwritten Chinese character recognition. Mostafa et al. [20] introduced OCFormer, a Transformer-based model, to improve the accuracy of optical character recognition (OCR) for Arabic handwritten texts. Furthermore, Campiotti et al. [21] presented an efficient OCR model that combines Convolutional Neural Networks (CNNs), Transformer encoders and Connectionist Temporal Classification (CTC) layers, demonstrating remarkable performance on the SROIE2019 dataset.

The success of ViT in character recognition is further exemplified by its application in various tasks. TrOCR model uses ViT as the encoder, combined with the pretraining text converter, to achieve advanced performance in the field of OCR [22]. Yang et al. [23] proposed the Transformer-based Radical Analysis Network (RTN), which improves the accuracy of Chinese character recognition. Rouhou et al. [24] introduced an end-to-end Transformer architecture that simultaneously achieves handwritten text recognition and named entity recognition.

Moreover, researchers have optimized ViT for specific character recognition challenges. Dan et al. [25] enhanced ViT with multi-level parallel branches to improve efficiency in Chinese character recognition. Geng et al. [26] designed LW ViT specifically for handwritten Chinese character recognition, combining Transformer with MobileNetV2, making it suitable for mobile deployment. Additionally, Azadbakht et al. [27] proposed the Multiplath ViT OCR model to tackle the license plate OCR problem.

Given the outstanding performance of ViT in character recognition, we plan to apply it to the more challenging task of cursive script recognition.

3 Method

3.1 Base models

We use two base models: TrOCR [22] and ViTSTR [28]. TrOCR uses the standard Transformer Encoder-Decoder structure. The TrOCR model contains two parts: an image transformer to capture features of images and a text transformer to generate wordpiece sequences. ViTSTR is a Vision Transformer-based model that, compared to ViT, can recognize multiple characters while ensuring the correct sequence and length of the characters. The only difference between ViT and ViTSTR is the prediction head, which is used to recognize multiple characters and their sequence.

3.2 Notation

The notations used in this section are presented in Table 1.

3.3 Training with Curriculum Learning

The concept of curriculum learning, derived from the educational system in human society [29], begins with easy concepts and gradually introduces more difficult ones to improve the efficiency and effectiveness of learning.

Typically, we start by filtering out the hard samples from the training set, focusing on a subset of easier samples for the initial stages of model training. After a certain number of training epochs, we gradually introduce subsets containing the more challenging samples. The two key challenges in this stage arise: distinguishing between easy and hard samples, and deciding when to incorporate the harder subsets into the training process. These methods are predefined, relying on prior human knowledge, such as manually annotated data sets, task-specific complexity metrics, and the popular scheduler *Baby Step* [30], which groups training data by difficulty and merges them after certain epochs. However, most current Chinese cursive scripts have no tag complexity metric for indicating the classification difficulty. Thus, we choose to deal with this problem during training process. To address these challenges, researchers decompose CL into two independent yet closely related subtasks [31], which is also the strategy we applied in our experiments. In this paper [32], these two subtasks are abstracted as **Difficulty Measurer** and **Training Scheduler**.

3.3.1 Difficulty Measurer

For the input feature x_i of each training sample i, the cross-entropy loss $\mathscr{L}(x_i)$ and confidence $\mathscr{C}(x_i)$ are computed by the classification model. The difficulty $\mathscr{D}(i)$ of each sample is then assessed using a combined metric.

$$\mathscr{D}(i) = \alpha \cdot \mathscr{L}(x_i) + \beta \cdot (1 - \mathscr{C}(x_i)) \tag{1}$$

Table 1: Notation used in this section.

Symbol	Description
i	A training sample
x_i	An input feature of training sample i
$\mathscr{L}(x_i)$	Cross-entropy loss for the sample i
$\mathscr{C}(x_i)$	Confidence predicted for the sample i
$\mathscr{D}(i)$	Difficulty metric for the sample i
τ	Threshold for classifying a sample as easy or hard
α	Weight coefficient for the cross-entropy loss
β	Weight coefficient for the confidence
ES	Set of easy samples
HS	Set of hard samples
L_{train}	The training loss
T_{pre}	Duration of the pretraining phase
T_{CL}	End time of the dynamic curriculum learning phase
N	Total number of samples in the training set
N_{easy}	Number of easy samples
p_j	The j-th flattened image patch
p_{class}	The class token
s_j	The j-th flattened skeleton patch
$\mathbf{E}$	Learnable embedding matrix for image patches
$\mathbf{E}_{\text{pos}}$	Positional embedding matrix
$\mathbf{E}_s$	Learnable embedding matrix for skeleton patches
$\|\|$	Concatenation operation along the embedding dimension
z_0	Combined embedding as input of ViT encoder

where α and β adjust the weights of the loss and confidence. The sample is classified as easy or hard based on whether$\mathscr{D}(i)$ exceeds a threshold τ .

$$\text{Difficulty}(i) = \begin{cases} \text{Hard Sample (HS)} & \text{if } \mathscr{D}(i) > \tau, \\ \text{Easy Sample (ES)} & \text{else if } \mathscr{D}(i) \leq \tau. \end{cases} \tag{2}$$

3.3.2 Training Scheduler

Initially, the entire training dataset is used for model training, which denoted as the pretraining phase T_{pre}. The training loss L_{train} at this stage includes all samples,

$$L_{\text{train}} = \frac{1}{N} \sum_{i=1}^{N} L(x_i), \quad \text{for } 0 \leq t \leq T_{\text{pre}}, \tag{3}$$

During the dynamic curriculum learning phase, from $T_{\text{pre}} + 1$ to $T_{\text{pre}} + n$, the training loss only

includes easy samples:

$$L_{\text{train}} = \frac{1}{N_{\text{easy}}} \sum_{i \in \text{ES}} L(x_i), \quad \text{for } T_{\text{pre}} + 1 \leq t \leq T_{\text{pre}} + n, \tag{4}$$

After this phase, from t>$T_{\text{pre}} + n$, both easy and hard samples are used:

$$L_{\text{train}} = \frac{1}{N} \left(\sum_{i \in \text{ES}} L(x_i) + \sum_{i \in \text{HS}} L(x_i) \right), \quad \text{for } t > T_{\text{pre}} + n \tag{5}$$

3.4 ViT with Skeleton Embedding

The Vision Transformer (ViT) [17] converts images into embeddings via patch-wise linear projection.

$$z_0 = [p_{\text{class}}; p_1\mathbf{E}; p_2\mathbf{E}; \ldots; p_N\mathbf{E}] + \mathbf{E}_{\text{pos}} \tag{6}$$

where z_0 is the input of ViT encoder, p_{class} is the class token, and $\mathbf{E}_{\text{pos}}$ is the positional embedding matrix.

In our modified ViT architecture, we integrate skeleton embeddings extracted from the images. For each sample, the embeddings from the image and its corresponding skeleton are concatenated before being fed into the ViT encoder. The combined embedding sequence is:

$$z_0 = [p_{\text{class}}; p_1\mathbf{E} \parallel s_1\mathbf{E}_s; p_2\mathbf{E} \parallel s_2\mathbf{E}_s; \ldots; p_N\mathbf{E} \parallel s_N\mathbf{E}_s] + \mathbf{E}_{\text{pos}} \tag{7}$$

where $p_j\mathbf{E}$ and $s_j\mathbf{E}_s$ represent the embeddings of the image and skeleton patches, respectively. The process is detailed in Figure 2.

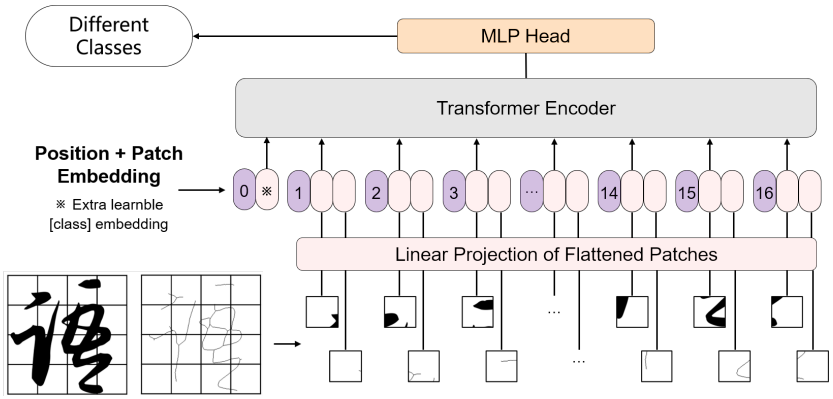

Fig. 2. Illustration of the modified Vision Transformer (ViT) architecture with skeleton embedding. The input image and skeleton embeddings are concatenated and combined with positional encoding before being fed into the encoder.

4 Experiment

4.1 Dataset

For the task of recognizing Chinese cursive calligraphy characters, the quality of the dataset is crucial [7]. Publicly available datasets for cursive calligraphy are limited, often derived from scanned images of ancient books, and typically require complex preprocessing or specialized model design.

To focus on the classification of cursive calligraphy images, we have developed a custom dataset consisting of single-character images labeled with their corresponding Chinese characters. The following sections will outline the generation, composition, and pre-processing steps of our dataset.

4.1.1 Dataset Generation

To simplify data pre-processing and achieve higher quality images of individual cursive characters, we chose to use TTF files for generation. We first carefully selected 20 noncommercial cursive font styles from some font library applications, obtaining their TTF files. Then we filtered the Unicode that were not within the dictionary range according to the label dictionary. After selecting the target glyphs, we rendered them into 512x512 pixel images and saved them to the dataset. The information of each image was recorded in a corresponding annotation file. The acquisition process is illustrated in the figure 3. Our dataset includes a total of 510 unique Chinese characters, each with nearly 20 samples, resulting in a total of 10,180 images.

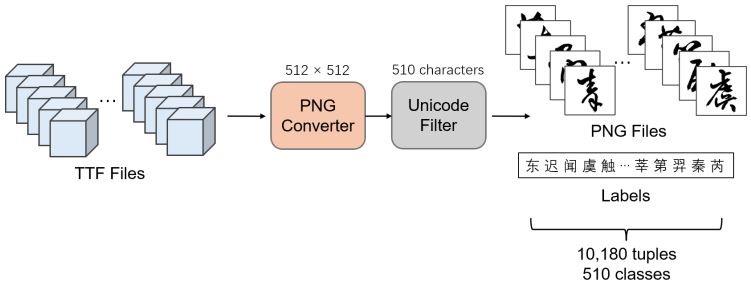

Fig. 3. The generation of our dataset. After converting TTF files into 512x512 images, the corresponding data pairs are obtained through filtering.

4.1.2 data augmentation

Given the artistic and aesthetic nature of cursive scripts, we draw inspiration from contrastive learning, highlighting the critical role of data augmentation in learning robust and discriminative features [33] [34]. Therefore, we emphasize the importance of data augmentation to enhance the

accuracy of Vi-T models in cursive script recognition. The parameters of the augmentation process are random translation using one of the following methods,

- Translated in the vertical and horizontal directions by a factor within the range of 0.1 to 0.2 of the image dimensions.
- Gaussian blur with a random kernel size randomly between 11 and 21.
- Gaussian noise with a standard deviation (std) randomly chosen between 0.3 and 0.5.

Each image underwent these three augmentation methods in a fixed sequence. We retained both the original images and the augmented images, thereby expanding the dataset and enhancing the model's ability to generalize across different styles and complexities of cursive handwriting.

4.1.3 skeletonization

Chinese characters are ideographic, with radicals and stroke structures rich in meaning. Cursive calligraphy not only reveals the structure of characters but also the style of strokes, such as thickness and connectedness, which can vary greatly among different authors, adding artistic and aesthetic value to the writing but is not essential for recognition tasks. Specifically, through structural information alone, we can accurately identify the specific unique Chinese character. Therefore, we hope to enhance Vi-T models for cursive calligraphy recognition task, drawing inspiration from current studies on skeletonization-based classification tasks for Chinese characters [21] and skeleton-based Chinese character generation tasks [35], though the present research in this area predominantly focuses on modern standard characters or font library characters rather than cursive calligraphy. Next, we performed skeleton extraction on each cursive calligraphy character image in our dataset, using this as a conceptually atypical form of prior knowledge to augment our dataset. For the cursive character images, we applied an iterative process of marking and removing edge pixels until only a single-pixel-wide skeleton remained. Throughout this process, we preserved the original character's geometric and topological structures (such as connectivity and holes). Examples of the skeleton extraction are illustrated in the following figure 4.

4.1.4 Test Dataset

Lacking an authoritative test dataset for our cursive single-character recognition, we randomly split the generated dataset into training, validation, and test subsets in an 8:1:1 ratio for subsequent use.

4.2 Settings

We use 4 NVIDIA GeForce RTX 3090 GPUs with the memory of 24GBs for training and 1 NVIDIA GeForce RTX 3090 GPU for testing.

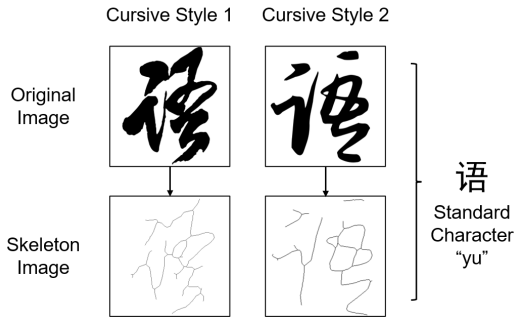

Fig. 4. The standard modern Chinese character "yu" in SimHei font (right), two cursive "yu" images from the dataset, and their corresponding one-pixel-wide skeleton images.

4.3 Dynamic Scheduling of Curriculum learning

When computing the difficulty $\mathscr{D}(i)$ of samples i, there are two cases of the value of α and β, which are the weight coefficients of cross-entropy loss and confidence scores,

- $\alpha = 1, \beta = 0$.
- $\alpha = 0, \beta = 1$.

4.3.1 About Cross-entropy Loss

For the initial 256 iterations, the dataset is dedicated entirely to the training process without exception. As the training progresses from batch 257 to batch 1000, a loss-based filtering criterion is introduced, where batches are discarded if their loss exceeds a threshold τ of 3. Subsequently, from batch 1001 up until the 2000th batch, this threshold τ is lowered to 1, tightening the criteria for batch inclusion. Beyond the 2000th batch, the threshold τ is further reduced to 0.1, allowing for a more inclusive approach where the vast majority of samples contribute to the training phase.

4.3.2 About Confidence scores

During the training process, the confidence score will continue to rise. We calculate each sample's confidence score, comparing it to the recent average confidence score, which resets every 1,024 samples. Samples with scores below average are masked, and if the number of masked samples in a batch exceeds the threshold, the batch will be skipped. The masking threshold also adjusts during training based on the recent average masked rate.

4.4 Results

4.4.1 Evaluation metrics

We evaluate the effectiveness of the model by the precision of test dataset, which is described as follow,

$$Precision = \frac{\text{Number of correct positive predictions}}{\text{Total number of positive predictions}}$$

4.4.2 Results of using Curriculum Learning

Table 2: Performance data for ViTSTR

BS	Sklt	CL	FE	Acc(%)	Prec	BS	Sklt	CL	FE	Acc(%)	Prec
32	-	-	-	82.809	0.810	64	-	-	-	82.809	0.879
			Resnet	83.006	0.845				Resnet	83.595	0.793
		Loss	-	**87.917**	**0.914**			Loss	-	87.328	0.897
			Resnet	87.819	0.897				Resnet	82.024	0.845
		Conf	-	86.248	0.793			Conf	-	84.479	0.879
			Resnet	80.354	0.776				Resnet	86.248	0.862
	Sklt	-	-	85.855	0.862		Sklt	-	-	82.809	0.845
			Resnet	82.711	0.879				Resnet	82.417	0.828
		Loss	-	86.051	0.828			Loss	-	83.890	0.897
			Resnet	80.157	0.879				Resnet	83.792	0.897
		Conf	-	81.827	0.879			Conf	-	84.086	0.793
			Resnet	81.336	0.759				Resnet	81.925	0.828

In table 2, BS stands for Batch Size, Sklt indicates using skeleton embedding, FE represents the feature extraction tool, CL denotes Course Learning, Conf is the confidence method, Acc stands for Accuracy, and prec represents Precision.

As demonstrated in Table 2, when the ViTSTR model is employed with a batch size of 32, both curriculum learning approaches—those grounded in the loss function and those based on confidence scores—substantially boost the model's predictive accuracy. However, when the batch size is increased to 64, employing the loss function for curriculum learning remains the most effective strategy, whereas reliance on confidence scores detrimental to the model's performance.

As figure 5 shown, for the TrOCR model, the adoption of curriculum learning strategies involving the Loss function or confidence scores does not yield a pronounced enhancement in the model's recognition accuracy.

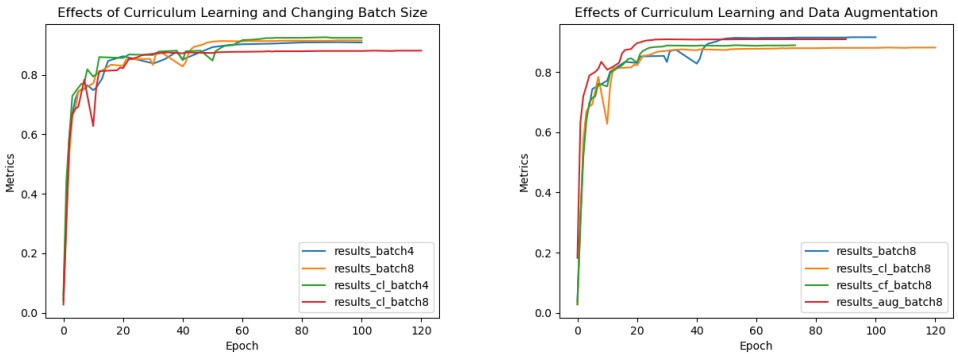

Fig. 5. Effects of Curriculum Learning and Changing Batch Size of TrOCR

4.4.3 Results of using Skeleton Embedding

The integration of the skeleton embedding has markedly enhanced the performance of the ViTSTR model, particularly evident when the batch size is set to 32. Conversely, when the batch size is doubled to 64, the performance improvement observed is less pronounced. Although CL and skeleton embedding alone can both enhance the performance, using CL method and skeleton embedding simultaneously is not effective.

4.4.4 Effects of Other Parameters and Strategies

Within a certain range, changes in batch size have a limited impact on model performance. The ViTSTR model achieves its best overall performance with a batch size of 32 samples, while the TrOCR model achieves higher accuracy with a batch size of 8 images. Using Resnet for feature extraction decreases the performance of the ViTSTR model. Data augmentation significantly improves the convergence rate of the model but has a limited effect on accuracy. As shown in the figures 5, changes in batch size have little impact on the convergence rate of the ViTSTR model.

5 Conclusion and Future work

This paper introduces a novel method for recognizing Chinese cursive calligraphy characters, utilizing Vision Transformer (ViT)-based models enhanced by curriculum learning and skeleton embeddings. The curriculum learning strategy allows the models to gradually increase the complexity of the training samples, leading to better convergence and more robust recognition capabilities. The incorporation of skeleton embeddings adds critical structural information, improving the model's ability to discern intricate cursive characters. Experimental results on Chinese cursive calligraphy datasets demonstrate significant improvements in recognition accuracy, underscoring the effectiveness of our approach. Future research could explore the integration of additional features and re-

finement of the curriculum learning strategy to further enhance model performance in this complex recognition task.

6 Acknowledgement

Xinrui Shan, Yilin Fang, Jinyang Zheng, and Tianhong Qi contributed equally to this work and should be considered co-first authors.

References

[1] Lu Shen, Bidong Chen, Jianjing Wei, Hui Xu, Su-Kit Tang, and Silvia Mirri. The challenges of recognizing offline handwritten chinese: A technical review. *Applied Sciences*, 13(6):3500, 2023.

[2] Cheng Lin Liu, Fei Yin, Da Han Wang, and Qiu Feng Wang. Online and offline handwritten chinese character recognition: Benchmarking on new databases. *Pattern Recognition*, 46(1):155–162, 2013.

[3] Sargur N. Srihari, Xuanshen Yang, and Gregory R. Ball. Offline chinese handwriting recognition: an assessment of current technology. *Frontiers of Computer Science in China*, 1(2):137–155, 2007.

[4] Wei Tang, Yijun Su, Xiang Li, Daren Zha, Weiyu Jiang, Neng Gao, and Ji Xiang. Cnn-based chinese character recognition with skeleton feature. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V*, page 461–472, Berlin, Heidelberg, 2018. Springer-Verlag.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Xiao Qin, Jianhui Jiang, Wei Fan, and Changan Yuan. Chinese cursive character detection method. *The Journal of Engineering*, 2020(2), 2020.

[7] Jung Liang, Wen-Hung Liao, and Yi-Chieh Wu. Toward automatic recognition of cursive chinese calligraphy : An open dataset for cursive chinese calligraphy text. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–5, 2020.

[8] Jingye Chen, Bin Li, and Xiangyang Xue. Zero-shot chinese character recognition with stroke-level decomposition. *CoRR*, abs/2106.11613, 2021.

[9] Huihui Si. Analysis of calligraphy chinese character recognition technology based on deep learning and computer-aided technology. *Soft Computing*, 28(1):721–736, 2024.

[10] Zhiyi Zhang, Lianwen Jin, Kai Ding, and Xue Gao. Character-sift: A novel feature for offline handwritten chinese character recognition. In *2009 10th International Conference on Document Analysis and Recognition*, pages 763–767, 2009.

[11] Xiao Qin, Jianhui Jiang, Wei Fan, and Changan Yuan. Chinese cursive character detection method. *The Journal of Engineering*, 2020, 07 2020.

[12] Zhao Zhong, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Handwritten chinese character recognition with spatial transformer and deep residual networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3440–3445, 2016.

[13] N Shanthi and K Duraiswamy. A novel svm-based handwritten tamil character recognition system. *Springer-Verlag*, 2010.

[14] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017.

[15] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Raymond Ptucha, Felipe Petroski Such, Suhas Pillai, Frank Brockler, Vatsala Singh, and Paul Hutkowski. Intelligent character recognition using fully convolutional neural networks. *Pattern Recognition*, 88:604–613, 2019.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[18] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. 2022.

[19] Ji Gan, Yuyan Chen, Boxia Hu, Jiaxu Leng, Weiqiang Wang, and Xinbo Gao. Characters as graphs: Interpretable handwritten chinese character recognition via pyramid graph transformer. *Pattern Recognit.*, 137:109317, 2023.

[20] Aly Mostafa, Omar Mohamed, Ali Ashraf, Ahmed Elbehery, Salma Jamal, Ghada Khoriba, and Amr S. Ghoneim. Ocformer: A transformer-based model for arabic handwritten text recognition. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 182–186, 2021.

[21] Israel Campiotti and Roberto Lotufo. Optical character recognition with transformers and ctc. In *Proceedings of the 22nd ACM Symposium on Document Engineering*, DocEng '22, New York, NY, USA, 2022. Association for Computing Machinery.

[22] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv e-prints*, 2021.

[23] Killian Barrere, Yann Soullard, Aurélie Lemaitre, and Bertrand Coüasnon. A light transformer-based architecture for handwritten text recognition. In Seiichi Uchida, Elisa Barney, and Véronique Eglin, editors, *Document Analysis Systems*, pages 275–290, Cham, 2022. Springer International Publishing.

[24] Ahmed Cheikh Rouhou, Marwa Dhiaf, Yousri Kessentini, and Sinda Ben Salem. Transformer-based approach for joint handwriting and named entity recognition in historical document. *Pattern Recognition Letters*, 155:128–134, 2022.

[25] Yongping Dan, Zongnan Zhu, Weishou Jin, Zhuo Li, and Mario Versaci. Pf-vit: Parallel and fast vision transformer for offline handwritten chinese character recognition. *Intell. Neuroscience*, 2022, jan 2022.

[26] Shiyong Geng, Zongnan Zhu, Zhida Wang, Yongping Dan, and Hengyi Li. Lw-vit: The lightweight vision transformer model applied in offline handwritten chinese character recognition. *Electronics*, 12(7), 2023.

[27] Alireza Azadbakht, Saeed Reza Kheradpisheh, and Hadi Farahani. Multipath vit ocr: A lightweight visual transformer-based license plate optical character recognition. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 092–095, 2022.

[28] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis, 2019.

[29] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[30] Valentin I. Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In Ron Kaplan, Jill Burstein, Mary Harper, and Gerald Penn, editors, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, Los Angeles, California, June 2010. Association for Computational Linguistics.

[31] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544. PMLR, 09–15 Jun 2019.

[32] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2022.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[34] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *CoRR*, abs/2004.11362, 2020.

[35] Yiming Gao and Jiangqin Wu. Gan-based unpaired chinese character image translation via skeleton transformation and stroke rendering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):646–653, Apr. 2020.