applied sciences

MDPI

Article

Stroke-Based Autoencoders: Self-Supervised Learners for Efficient Zero-Shot Chinese Character Recognition

Zongze Chen [1], Wenxia Yang [1,*] and Xin Li [2]

[1] Department of Mathematics, Wuhan University of Technology, Wuhan 430070, China
[2] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, VA 26506, USA
* Correspondence: wenxiayang@whut.edu.cn

Abstract: Chinese characters carry a wealth of morphological and semantic information; therefore, zero-shot Chinese character recognition with the morphology of Chinese characters has drawn significant attention. The previous methods are mainly based on radical-level decomposition or stroke-level decomposition, which usually cannot capture adequately the structural and spatial information of Chinese characters. In this paper, we develop a stroke-based autoencoder (SAE), to model the sophisticated morphology of Chinese characters with a self-supervised method. Following its canonical writing order, we first represent a Chinese character as a series of stroke images with a fixed writing order, and then our SAE model is trained to reconstruct this stroke image sequence. This pre-trained SAE model can predict the stroke image series for unseen characters, as long as their strokes or radicals are in the training set. We have designed two contrasting SAE architectures on different forms of stroke images. One is fine-tuned on existing stroke-based method for zero-shot recognition of handwritten Chinese characters, and the other is applied to enrich the Chinese word embeddings from their morphological features. The experimental results validate that after pre-training, our SAE architecture outperforms other existing methods in zero-shot recognition and enhances the representation of Chinese characters with their abundant morphological and semantic information.

Keywords: Chinese character; self-supervised learning; zero-shot recognition; pre-training model; semantic enhancement

check for updates

Citation: Chen, Z.; Yang, W.; Li, X. Stroke-Based Autoencoders: Self-Supervised Learners for Efficient Zero-Shot Chinese Character Recognition. *Appl. Sci.* **2023**, *13*, 1750. https://doi.org/10.3390/app13031750

Academic Editors: Deng Chen and Hong Zheng

Received: 31 December 2022
Revised: 20 January 2023
Accepted: 24 January 2023
Published: 30 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Chinese characters emerged as hieroglyphs and naturally carry very rich morphological and semantic information. Therefore, the pictographic information of Chinese characters can be applied to explore the application of Chinese character structures, such as zero-shot Chinese character recognition [1,2] and the generation of stylish Chinese fonts [3,4]. The information can also facilitate many language-dependent NLP tasks, such as the recognition of Chinese names [5,6] and the segmentation of Chinese words [7]. Zero-Shot Chinese character recognition (zero-shot CCR), as one application of Chinese character structures, has been involved in many real-world applications [2], such as handwritten signature identification [8,9], and historical documents retrieval [10,11]. The main idea of zero-shot CCR is to recognize the characters that have not appeared in the training set by learning the infrastructure of Chinese characters between unseen classes and seen classes. This infrastructure can be explored from radical-level [2,12–17] or stroke-level [1,2].

Radical-based methods learn the infrastructure of Chinese characters by dividing them into radical sequences, and boost significant performance on zero-shot recognition. However, due to the large number of radical parts in Chinese characters, the distribution of different radical parts is usually uneven thus may cause serious class imbalance problem [1]. To tackle this challenge, Chen et al. [1] used stroke-based encoding by dividing strokes into

five basic categories, making breakthrough progress in zero-shot recognition. However, this stroke-based approach still has some limitations. Specifically: (i) It uses manual decomposition to label 2D image information into a series of numbers, which cannot take full advantage of the spatial information of Chinese characters. (ii) It depends heavily on the number of classified strokes, and the strokes in Chinese characters are often to be divided into twenty-eight [3] or thirty-eight [18] categories. Too many classifications may lead to imbalance problems, while too few classifications may lead to repetition [1]. Finding the best number of categories for manual decomposition of Chinese character is a time-consuming and laborious work. In short, how to find a suitable label to describe the structural information of Chinese characters is still an opening challenge.

From another perspective, the reliance on labeled data has always been a major obstacle to the development of deep learning. A good labeled image database, such as ImageNet [19], is instrumental in advancing computer vision and deep learning research. Recently, this appetite for labeled data is successfully addressed in natural language processing and computer vision by self-supervised learning. BERT [20] and MAE [21], with pre-trained methods, have achieved great success on related tasks. Thus, we expect to explore a self-supervised method on Chinese character images to avoid manual labeling, which facilitates the extraction of the structural and spatial information of Chinese character images in a more comprehensive style.

Self-supervised learning on images learns image representations by recovering masked images based on the encoding of input images. Inspired by Transformer, Dosovitskiy et al. [22] apply Transformer directly to the sequences of image patches for image recognition, and Bao et al. [23] propose a masked image modeling task to pre-train Vision Transformers. Most recently, masked autoencoder (MAE) [21] masks a high proportion (75%) of the input images and performs well on downstream tasks such as image recognition and instance segmentation. However, these methods cannot be applied directly to Chinese character images, since there are great differences between Chinese character images and general images in terms of both image structure features and semantic information. To be specific, **(i)** Chinese character images are usually space discrete, with the smallest unit stroke, while the general images are continuous with unit pixel, so the random mask may not work; **(ii)** Chinese characters are human-generated signals that are information dense with highly structured semantics, and every stroke or radical part may have its special informative meaning; **(iii)** Each Chinese character can be represented as a time series of images that carry information throughout its fixed order of strokes, as shown in Figure 1.

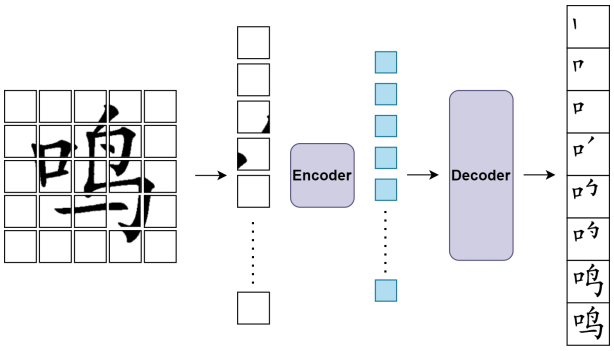


Figure 1. Our SAE architecture based on Vision Transformer. The encoder is applied on the image patches of the Chinese characters (we take '鸣' for example). The decoder reconstructs the stroke sequence images (the order of strokes) from the encoded patches.

Driven by the above analysis, we propose a stroke-based autoencoder (SAE) to extract the morphological information from the sequences of each Chinese character images. Our framework consists of two parts: **(i)** the encoder network, which operates on the image patches of Chinese characters to obtain encoded patches, and **(ii)** the decoder network, which reconstructs the stroke sequence images of each corresponding Chinese character.

The encoder in our SAE pre-training model aims at extracting the structured information from stroke images, and the decoder learns to predict the sequences of the writing order of Chinese characters. Some example results are illustrated in Figure 2. We choose some Chinese characters, all of which have eight strokes, for a neat illustration. The left column (with a yellow background) is the images of Chinese characters '鸣', '图', '茄' and '迢' in training dataset, and the right column (with a pale blue background) are images of '莺', '咚', '迦' and '苕' in the validation data set. Each Chinese character can be divided into two radical components, as illustrated in the middle column (with a light green background). These characters are made up of seven basic radical components, as illustrated at the bottom of the figure. For each character, the first row is the original stroke sequence images, i.e., their canonical order of strokes, and the second row is the corresponding stroke sequence predicted by our SAE model. The example results demonstrate that our pre-training model can successfully predict and reconstruct the unseen strokes and radicals of Chinese characters.

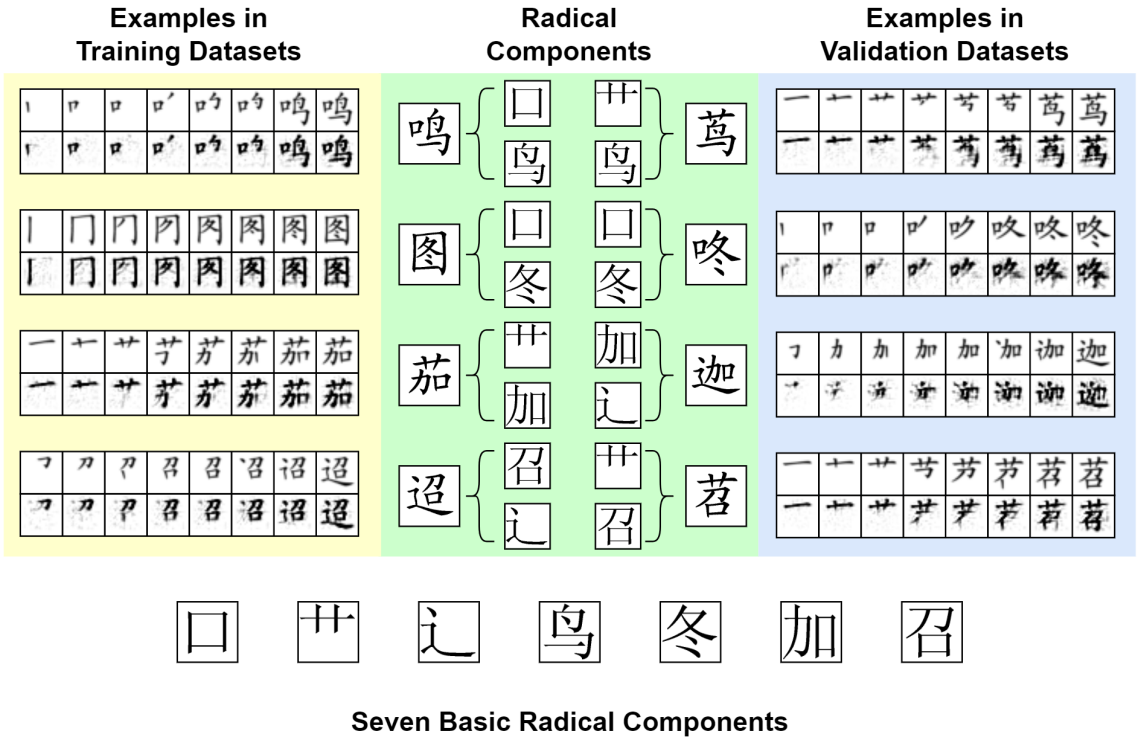


Figure 2. Example results on training (seen class) and validation (unseen class) datasets. Each character has the original picture (above) and reconstructed picture (below). The left column is the Chinese character image in the training datasets, and the right column is the Chinese character image in the validation datasets. The four characters of each column consist of seven basic radical parts.

Therefore, our SAE fully learned the morphological information of Chinese characters, which can be applied to explore the application of Chinese character structures, such as zero-shot Chinese character recognition and the generation of stylish Chinese fonts. It can also be applied to enrich word embeddings for downstream NLP tasks such as Chinese Named Entity Recognition. In this paper, we focus on the morphological structures of Chinese characters and utilize our pre-training models to explore zero-shot Chinese character recognition and word embeddings. we design two different SAE architectures on different forms of stroke images for different tasks. To the best of our knowledge, we are the first one to systematically explore the stroke sequence restoration and prediction at the image level. The experimental results validate that our SAE outperforms the previous results [1,12,13] on zero-shot recognition of Chinese characters. What is more encouraging is that our SAE is pre-trained on a small and concise dataset with printed Chinese character images,

but boosts huge performance in zero-shot recognition on a large and complicated dataset consisting of handwritten Chinese character images. Meanwhile, the word embedding generated by our SAE is used to calculate the word similarity of Chinese characters and their radical parts as an intrinsic task. The experiment shows that our SAE can reliably extract rich semantic information from Chinese characters.

The main contributions of our work are as follows:

(1) We initially present a stroke-based autoencoder (SAE) pre-training model through self-supervised learning. The pre-trained SAE model successfully captures the morphological and semantic information of Chinese characters.

(2) We pre-train SAE on a small and concise dataset with printed Chinese character images, but boosts huge performance on a large and complicated dataset consisting of handwritten Chinese character images.

(3) The pre-trained SAE model is fine-tuned with the existing stroke-based method and outperforms other existing methods in unseen (zero-shot) and seen Chinese character recognition. This model can also be applied to enrich Chinese word embeddings from their morphological structures.

2. Related Work

Masked Image Encoding: To learn image representations by predicting masked images from an input sequence of images, DAE [24] first considers masks as noises and obtain image representation by denosing the corrupted input signals. This pioneering work is generalized to a number of masked image encoding methods. Inspired by the Transformer, many approaches [22,23,25] are proposed for image encoding. The Vision Transformer (ViT) [22] performs a preliminary exploration of masked patch prediction for self-supervised learning. Beit [23] predicts the discrete tokens [26] to pre-train ViT. Most recently, MAE [21] masks about 75% proportion of the input image and achieves better recognition accuracy on ImageNet-1K and other datasets. However, Chinese character images are quite different from general images in terms of both image structure features and semantic information. Thus, we explore a novel self-supervision method that focuses on the morphological structures of Chinese characters by reconstructing their stroke image sequences.

Zero-Shot Chinese Character Recognition: Zero-Shot Chinese character recognition depends on the set of labeled training set of seen classes and some preliminary knowledge about the semantic relation between unseen classes and seen classes [27]. This semantic relation is often described as a high-dimensional vector, which is established from a semantic embedding space (e.g., the attribute space [28]). Specifically, most existing zero-shot learning methods learn a projection function from the visual feature space to the semantic embedding space [27], by training visual data with labels consisting only of visible classes. For zero-shot Chinese character recognition, the semantic embedding space is heavily based on the exploration of the structures of Chinese characters, which can be explored from radical-level [2,12–17] and stroke-level [1,2]. These radical-based methods establish the semantic relation by dividing Chinese characters into radical sequences. However, some radicals may not appear in training sets in a data hungry condition, namely the class imbalance problem [1]. Thus, a more refined decomposition of Chinese characters' structure remains to explore.

Cw2vec [29] first exploits the stroke-level information of Chinese characters with stroke n-grams, which divides Chinese characters into five basic categories of strokes such as horizontal, vertical, left-falling, right-falling, and turning, according to the Chinese national standard GB18030-20, as shown in Figure 3. For example, the Chinese character '鸣' can be encoded as '25135451' as shown in Figure 4. Chen et al. [1] then used this stroke-based encoding as the semantic embedding of the Chinese character, making breakthrough progress in zero-shot recognition. However, this method depends heavily on the number of classified strokes and usually fails to capture adequate spatial and morphological

information about Chinese characters. To fix this problem, we use the self-supervised method to explore the structure of Chinese characters without these manual decomposition.

Category	Horizontal	Vertical	Left-falling	Right-falling	Turning
Shape, ID	一(㇀), 1	丨(亅), 2	丿, 3	丶(㇏), 4	乛(㇕乚), 5

Figure 3. Five basic categories of strokes.

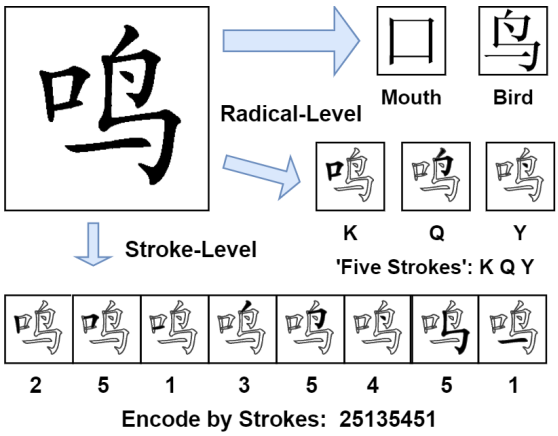


Figure 4. Encoding methods at radical-level and stroke-level

Semantic Enhancement from the morphology of Chinese Characters: Cao et al. [29] first exploit the stroke-level information of Chinese characters with stroke n-grams. Li et al. [6] encode Chinese characters with Five-Stroke for the recognition of Chinese Named Entities. However, these human-generated encoding methods may lead to repetition, since they cannot take full advantage of the spatial information of Chinese characters. Instead of using the ID-based character embedding method, Sun et al. [7] and Li et al. [30] extract the glyph information of Chinese characters for the language model based on Chinese character images. However, due to the lack of supervision, these methods cannot focus on global and local information simultaneously. Thus, we attempt to use self-supervised methods to extract their semantic information by reconstructing their stroke image sequences.

In addition, word embedding evaluation plays a important role in the estimation of semantic enhancement. Evaluation methods can be roughly divided into two categories: Intrinsic evaluation with word similarity [31] or word analogy [32], and extrinsic evaluation for the performance of downstream NLP tasks. It has been verified that there are strong correlations between intrinsic and extrinsic evaluations [33].

3. Prior Structural Knowledge about Chinese Characters

In this section, we provide some prior knowledge of the structures of Chinese characters to help us better understand Chinese characters.

Chinese characters are hieroglyphs, which can be divided into radical level or stroke level [1]. At the radical level, Chinese characters with similar structures often have similar meanings. As shown in Figure 4, the meaning of the Chinese character '鸣' is 'the chirping of birds', and '鸣' can be split into two radicals: '口' (mouth) and '鸟' (bird). Thus, Chinese characters are naturally informative datasets [34,35]. This abundant information is often used to enrich Chinese word embeddings in NLP tasks [29]. Meanwhile, according to the Five Strokes input method, most Chinese characters can be divided into no more than four basic Chinese character structures. Based on the 'Five Strokes input method', GlyphCRM [6] provides a bidirectional encoding method for Chinese characters with the Glyph. As shown in Figure 4, '鸣' also can be encoded as 'KQY' using the 'Five Strokes input method'.

At stroke level, Chinese characters can be divided into five basic categories of strokes [29]: horizontal, vertical, left falling, right falling, and turning according to the Chinese national standard GB18030-2005, as shown in Figure 3. Compared to radical level, the stroke-level based presentation has superiority of conciseness and consistency, while radical level is more complex with less space effectiveness. Moreover, Chinese characters have some general writing principles such as "from left to right", "top to bottom", "inside to outside". With this principle, '鸣' can be encoded by strokes as '25135451', as illustrated in Figure 4. This property reminds us that when the model learns the orders of strokes of some characters, it can naturally master how to write new characters, even if these characters have never appeared in the training set.

4. Approach

In this section, we introduce the proposed SAE architectures and the pre-training model. We develop two contrasting architectures for different tasks. For zero-shot Chinese character recognition, we propose an SAE-based on Resnet and Transformer, while for semantic enhancement, we propose an SAE-based on Vision Transformer. In the experimental part, we will explain why these two models match with different Chinese NLP tasks.

4.1. The Pre-Training Model

Our stroke-based autoencoder (SAE) is a self-supervised model that predicts the original signal $\hat{z} = (z_1, \ldots, z_n), z \in \mathbb{R}^{n \times d_3}$ given its encoded patches, as shown in Figures 1 and 5. Similar to other autoencoders, our model has an encoder that maps an input sequence of symbol representations $x = (x_1, \ldots, x_n), x \in \mathbb{R}^{n \times d_1}$ to a sequence of continuous representations $y = (y_1, \ldots, y_n) \in \mathbb{R}^{n \times d_2}$, and a decoder that generates an output sequence $z = (z_1, \ldots, z_n) \in \mathbb{R}^{n \times d_3}$, where d_1 is the number of pixels in the input patches, d_3 is the dimension of the outputs and d_2 is the dimension embedded of the encoded patches. Unlike the Masked Autoencoder [21], which uses the spatial segmentation of pictures, we divide Chinese pictures by time according to their writing orders, since every Chinese character has its unique writing order. In this way, we reconstruct the stroke sequence of character images from the encoded patches.

SAE based on Vision Transformer: Our encoder based on Vision Transformer(ViT) [22] is applied to the image patches of a Chinese character picture, as shown in Figure 4. For a Chinese character, we set its stroke number to m. In our Chinese character datasets, the maximum stroke number of characters is 24. Thus, we split each image into $n = 5 \times 5$ patches $x = (x_1, \ldots, x_n)$. For these characters whose stroke number is less than 25, we supplement their stroke sequence images with blank pictures. We apply a linear projection(Proj) with added positional embeddings(Pos) on the input patches; after that, a series of Transformer blocks are used to process the encoded patches. The encoder has depth 4 and width 256 of Transformer blocks.

$$y = Transformer_Blocks_1(Proj(x) + Pos) \tag{1}$$

$$y = Transformer_Blocks_i(y), i = 2, 3, 4 \tag{2}$$

Our decoder is similar to the decoder of Transformer [36] applied on encoded patches, but only consists of multihead self-attention. That is, we directly reconstruct the image sequence of a Chinese character with encoded patches. The decoder also has depth 4 and width 256 of the Transformer blocks.

$$z = Transformer_Blocks_1(y + Pos) \tag{3}$$

$$z = Transformer_Blocks_j(z)), j = 2, 3, 4 \tag{4}$$

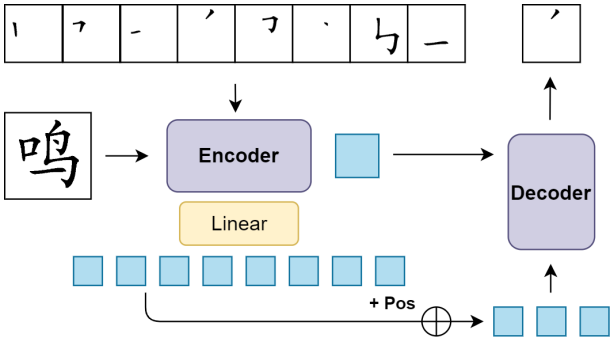


Figure 5. Our SAE architecture based on Resnet and Transformer. The encoder is applied on the images of Chinese characters and their image sequence of labels. The decoder FL reconstructs the stroke sequence images through the encoded patches of original image and stroke sequence images at the previous moment.

SAE based on Resnet and Transformer: If we consider the stroke sequence of images as a sentence, each image can be viewed as a word. Thus, the reconstruction task can be modeled as a sequence-to-sequence (seq2seq) task. In this regard, we exploit the Transformer [36] to decode the stroke sequence of images from the previous images and the original image of a Chinese character. Considering that Resnet [37] plays an important role in recognition tasks and performs better on smaller datasets than Vision Transformer, we propose an SAE architecture based on Resnet and Transformer, the overall framework is illustrated in Figure 5. For an input image $x_m \in \mathbb{R}^{C\times H\times W}$, we employ Resnet as the encoder to output $\hat{y} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 256}$ encoded patches. For input image sequences x, a linear projection(Proj) is added after Resnet to map $n \times \frac{H}{2} \times \frac{W}{2} \times 256$ to output $y \in \mathbb{R}^{n\times 256}$.

$$\hat{y} = Encoder(x_m) \tag{5}$$
$$y = Proj(Encoder(x)) \tag{6}$$

The encoder here has the effect of generating word vectors from image sequences. The decoder has depth 1 and width 256 of Transformer blocks. The detailed architecture of our encoder and decoder is shown in Figure 6. We apply a masked multi-head attention (*Masked_MultiHead*) with added positional embeddings(Pos) on input sequence y. Then, a layer normalization [38] is employed after residual connection.

$$y^{(1)} = y + Pos \tag{7}$$
$$y^{(2)} = Masked_MultiHead(y^{(1)}, y^{(1)}, y^{(1)}) \tag{8}$$
$$y^{(3)} = LayerNorm(y^{(1)} + y^{(2)}) \tag{9}$$

Next, we employ multi-head attention (MultiHead) on $\hat{y}$ and $y^{(3)}$. A full connection layer(FC) is added to output z at last after layer normalization and residual connection.

$$y^{(4)} = MultiHead(\hat{y}, y^{(3)}, y^{(3)}) \tag{10}$$
$$y^{(5)} = LayerNorm(y^{(3)} + y^{(4)}) \tag{11}$$
$$z = FC(y^{(5)}) \tag{12}$$

Reconstruction target:. In the pre-training stage, our SAE reconstructs the image by predicting the pixel values of the encoded patches. The last layer of the decoder is a full connection layer whose number of output channels is equal to the number of pixels in each patch. The output of the decoder is then reshaped into a reconstructed image.

The reconstruction loss function computes the mean squared error(MSE) between the reconstructed and original images in pixel space as as Equation (13):

$$loss = \frac{1}{n}\sum_{i}^{n}(z_i - \hat{z}_i)^2 \tag{13}$$

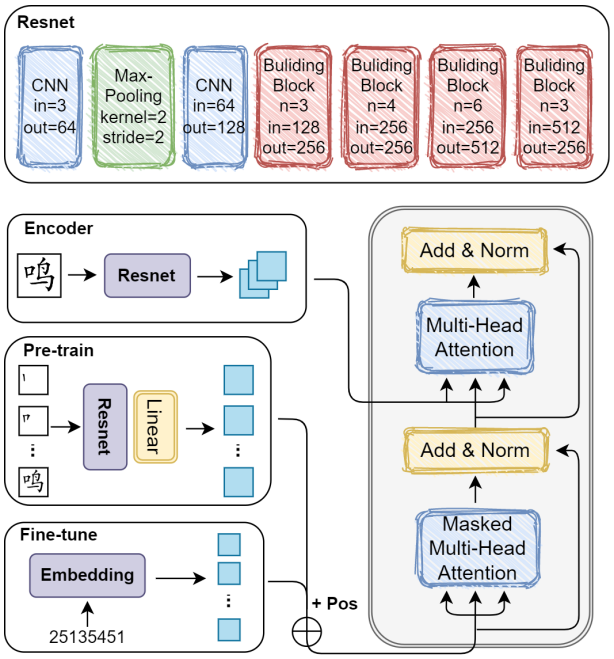


Figure 6. The detailed architecture of encoder and decoder.

4.2. Fine-Tuning for Zero-Shot Chinese Characters Recognition

Chen et al. [1] propose a stroke-based method for zero-shot Chinese character recognition, which makes breakthrough progress in zero-shot recognition. This method uses ResNet [37] as the Image-to-Feature Encoder and uses a series of Transformer blocks [36] as the Feature-to-Stroke Decoder. We fine-tune our pre-trained model on this method for zero-shot Chinese character recognition. The basic architecture is similar to the stroke-based autoencoders illustrated in Figure 5. The ground-truth is denoted by $t = (t_1, t_2, \ldots, t_n)$, and the Cross-Entropy loss is used to optimize the model as Equation (14):

$$loss = -\sum_{i}^{n} log(p(t_i)) \tag{14}$$

where $p(t_i)$ is the probability of class t_i at the $i-th$ time step. Besides, we notice that different Chinese characters may have the same stroke sequences, that is, one label may correspond to multiple Chinese characters. As Figure 7 illustrates, the four characters '甲', '叶', '叮' and '申' have the same stroke sequences. To tackle this one-to-many problem, Chen constructs the confusable set C, which includes these characters that have the same stroke sequence, and then feeds both the input handwritten character image and the printed images in set C into the encoder. After getting the encoded patches of the input image(denoted by F) and the encoded patches of printed images in C (denoted by $\mathcal{F}^{'} = \{F_1^{'}, F_2^{'}, \ldots, F_N^{'}\}$), the similarity between F and $F_i^{'}$ is calculated, and the final Chinese character is determined by Equation (15).

$$i^* = \underset{i \in \{1,2,\ldots,N\}}{\arg\max} \frac{F^T F_i^{'}}{\|F\| \times \|F_i^{'}\|} \tag{15}$$

For our zero-shot recognition task, we apply the SAE based on ResNet and Transformer to predict the stroke sequences of Chinese characters after pre-training. The last layer of the decoder is a linear projection whose number of output channels changes from the number of pixel values to the number of categories. The loss function for fine-tuning computes the cross entropy loss between the prediction and the ground truth.

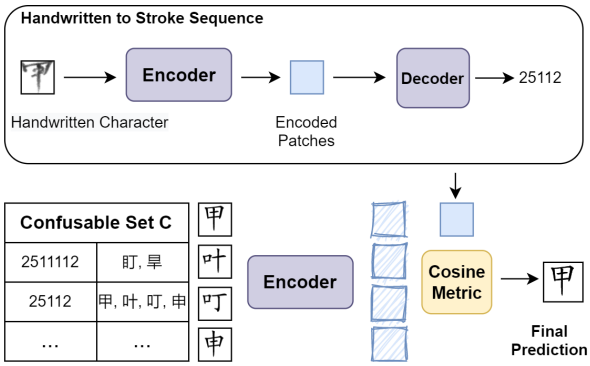


Figure 7. The architecture of the stroke-based zero-shot recognition method.

For a general transfer learning task, it is essential to pre-train the model on a large dataset to obtain the abundant knowledge. Then, the pre-trained model can generalize well to the target task with a limited dataset. For our zero-shot tasks, however, pre-training datasets are much smaller than fine-tuning datasets; in addition, the handwritten Chinese characters images from fine-tuning datasets are more complicated than printed Chinese character images in the pre-training datasets. As a result, the model parameters learned by pre-training may not generalize well in our zero-shot tasks.

To fix this problem, we need to figure out the respective strengths of image reconstruction and zero-shot recognition, and explore how to combine them. Our SAE architecture has a better generalization performance in stroke sequence prediction by reconstructing the stroke sequence of images. This performance is learned from the deep layer of the encoder and decoder. The zero-shot recognition task teaches the model to extract rich and more generalized features from large and complicated datasets, i.e., handwritten Chinese character images. This performance is learned from the shallow layer of the encoder.

We present a new fine-tuning method by combining the training parameters learned from handwritten Chinese character images and pre-training parameters learned from printed Chinese character images. We first train the stroke-based model [1] on zero-shot recognition tasks, and overwrite the parameters of the last building block of the encoder along with the parameters of the decoder with our pre-trained parameters. Then, we freeze the parameters of the first three building blocks of the encoder, namely the parameters learned from handwritten Chinese character images. The rest of the parameters are fine-tuned for zero-shot tasks in the end.

5. Experiments

To evaluate the effectiveness of our proposed SAE model, we carry out comprehensive experiments on pre-training, unseen (zero-shot) Chinese character recognition and seen recognition. We demonstrate both qualitative and quantitative evaluations of the performance of our approach.

5.1. Datasets

For the pre-training procedure on recognition of unseen Chinese characters, we provide two forms of stroke sequence of printed Chinese characters. As shown in Figure 8, for the set of sequences of stroke images in the first form, each subsequent stroke image contains the previous stroke information of a Chinese character. For the second form, each subsequent stroke image contains only the individual strokes. For each form, we collected 36,670 stroke sequence images of 3755 printed Chinese characters from Hanzi Writer [39].

Each stroke corresponds to an image. For the reconstruction task, the first 500, 1000, 1500, 2000 and 2755 classes of 3755 printed Chinese character images are used as the training sets, and the last 1000 classes are used as the validation set. Furthermore, as marked in Figure 8, the third stroke image is a Chinese character '口', which contains the stroke information of the first three images. Thus, we can take the third stroke image as the input of SAE and predict the first three stroke images. We augment the datasets by taking the k-th stroke images to predict previous stroke images, where $k \in \{1, 2, \ldots, m\}$. In this way, we can improve the generalization performance of our pre-training model.

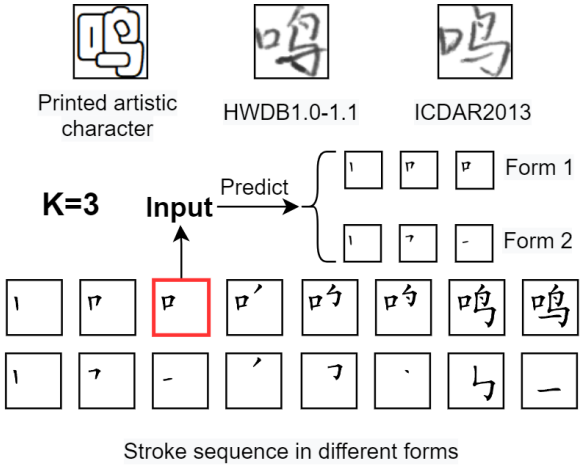


Figure 8. Examples of datasets. The last two lines of images are different forms of Chinese stroke sequences.

For the fine-tuning procedure on recognition of unseen Chinese characters, we use HWDB1.0-1.1 [40] as the training set, which contains 2,678,424 offline handwritten Chinese character images with 3881 classes collected from 720 writers. We use ICDAR2013 [41] as the validation set to evaluate the performance of our model. There are totally 224,419 offline handwritten Chinese character images and 3755 classes collected from 60 writers in ICDAR2013. We also use the 105 printed artistic fonts for 3755 characters (394,275 samples) collected by Chen et al. [1] for this recognition task. For the zero-shot recognition task, we choose samples from the first 500, 1000, 1500, 2000 and 2755 classes as the training set and the last 1000 classes as the validation set.

5.2. Implementation Details

Our model is implemented with PyTorch and the experiments are carried out on a server with 4 NVIDIA RTX2080 GPU with 8GB memory. AdamW [42] optimizer is used with the initial learning rate set to 1e-4. In addition, the cosine annealing rule [43]: $lr_t = \eta_m in + (lr_m ax - lr_m in) * (1 + cos\frac{T_{cur}}{T}\pi)$ is used to decay the learning rate, where T_{cur} and T represent the current epoch and the total number of epochs, respectively. The batch size is set be 32. The optimizer momentums β_1 and β_2 are set to be 0.9 and 0.999, the weight decay is set to be 0.001 empirically.

For ViT-based pre-training in SAE, the input images are resized to 140×140 and the label images are resized to 28×28. The input images of SAE based on Resnet and Transformer are resized to 28×28. Grayscale images are normalized to [0, 1].

Data augmentation is vital for pre-training on small datasets. As shown in Figure 2, for the characters '图' and '咚', they are both composed of the same size of '冬' and different size of '口'. Thus, it is important to generalize across glyphs of different sizes in a small dataset. We use random resized cropping by setting a random area scale between 0.8 and 1, and a random aspect ratio between $\frac{3}{4}$ and $\frac{4}{3}$ on the same set of samples and labels. Figure 9 shows the validation loss in different data sets with and without data augmentation. It can be seen that the validation loss has been dramatically reduced by data augmentation. For

fine-tuning in our zero-shot task, no other data augmentation strategies are used, since our dataset is large enough.

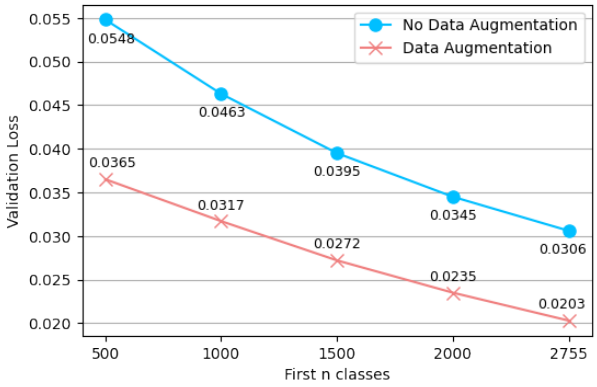


Figure 9. Validation loss in different data sets and data augmentation. The x-axis is the number of sample classes in the training dataset, i.e., first 500, 1000, 1500, 2000, 2755 classes. The y-axis is the MSE loss on last 1000 classes.

5.3. Experiments upon Pre-Training

In Section 4, we propose two contrasting architectures in different forms of stroke sequence images for different tasks. To compare the performance of the two architectures on different forms of data sets, the first 500, 1000, 1500, 2000 and 2755 classes of 3755 printed Chinese characters are used as the training sets, and the last 1000 classes are used as the validation set.

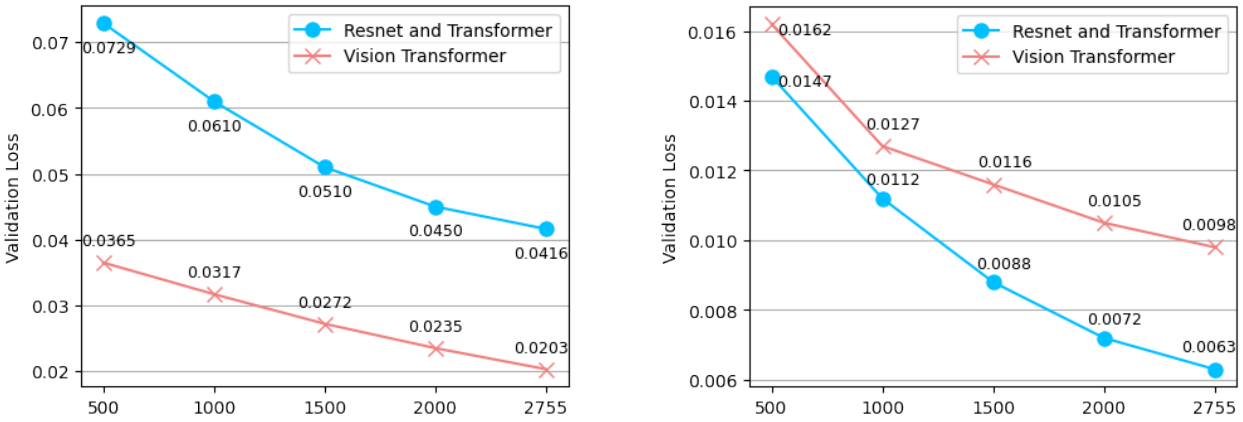
The validation loss in different architectures and different forms of data sets is illustrated in Figure 10. Figure 10a shows the validation loss of two architectures on the first form of the datasets, and Figure 10b shows the validation loss of two architectures on the second form of data sets. First, we can see that the validation loss on the first form of data sets is significantly higher than on the second form. This is because each image in the second form only contains a single stroke and most of the pixels are zero. Information is much more sparse than in the first form. Therefore, it is easier to reconstruct. Second, our SAE based on Resnet and Transformer predicts the next signal with teacher forcing strategy and, therefore, achieves smaller validation loss on second form of data sets. But for the first form of data sets, the former performs worse than the latter. This is because we tackle the reconstruction problem as a seq2seq task. We take the stroke sequence of images as a sentence and each image as a word. Since for the t-th image in the first form, it contains the preceding t-1 stroke information, which means that the previous images are redundant. Thus, when predicting the next signal, the SAE architecture based on Resnet and Transformer will only focus on the last image and ignore other images, which is inconsistent with the idea of a seq2seq task. As a result, this architecture generalizes significantly worse than the SAE based on ViT.

Therefore, we conclude that these two architectures are suitable for different forms of datasets and different tasks. For zero-shot Chinese character recognition task, we adopt the SAE based on ResNet and Transformer pre-trained on the second form of data sets. While for the semantic enhancement task, it is more efficient for us to adopt the SAE based on ViT, since it can predict the image sequences dependent on only one input image without teacher forcing strategy.

Moreover, the first form of stroke image reflects the overall structural changes of Chinese character strokes, it has more global information than the second form and help our model utilize the global and local information of Chinese characters simultaneously.

We pre-train our SAE (Vision Transformer with the first form of datasets) on 3755 Chinese Characters, and some example results has been illustrated in Figure 2. Figure 11 gives more results pre-trained on the second form of datasets. The experiment demon-

strates that after pre-training, our model can efficiently reconstruct the stroke sequence of unseen Chinese characters. This is because the encoder is capable of extracting semantic information from stroke images, and the decoder acquires the ability to predict the stroke sequences of Chinese characters. Thus, our SAE can be applied not only to enrich word embeddings, but can also explore the structures of Chinese characters in zero-shot Chinese character recognition.



(**a**) The first form of Chinese character stroke images. (**b**) The second form of Chinese character stroke images.

Figure 10. Validation loss on different forms of image sequence label and different architecture. (**a**) is the validation loss on the first form, (**b**) is the validation loss on the second form.

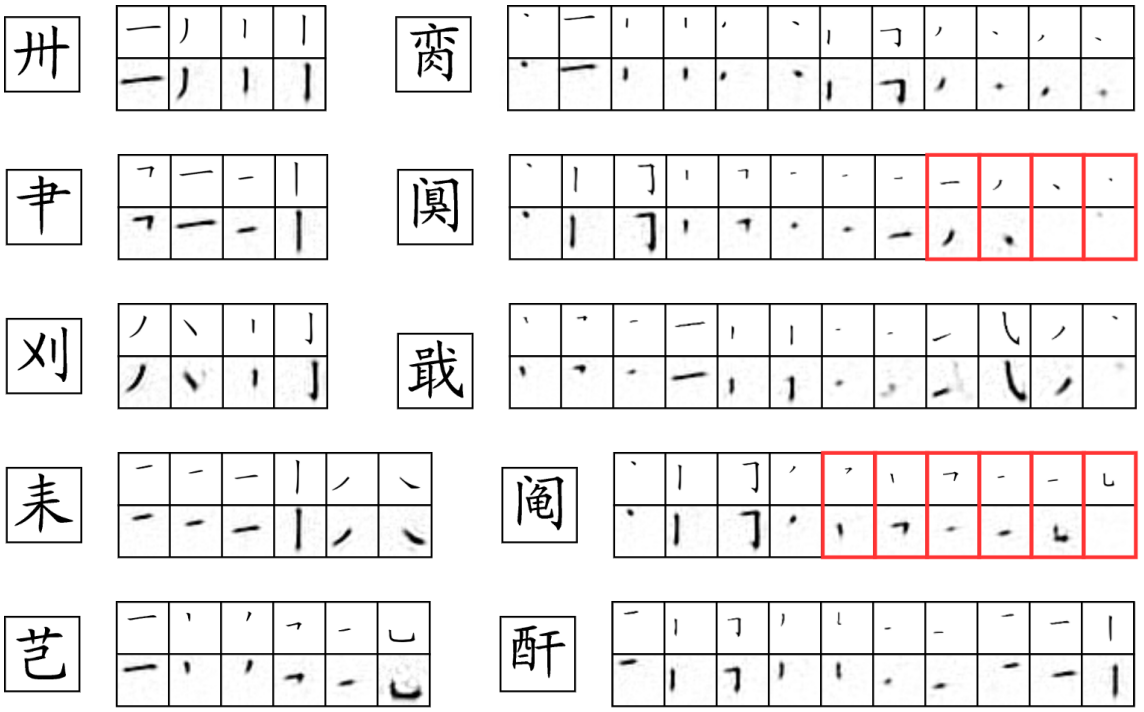


Figure 11. Example results on validation datasets of second form pre-trained on SAE based on Resnet and Transformer. Each character has the original picture (above) and reconstructed picture (below). Those enclosed in red boxes represent some wrong reconstructions.

Additionally, we notice that for the characters '阒' and '阄', there are some wrong reconstructions (highlighted in red boxes). This is because that both '阒' and '阄' are with semi-surrounded structures, and for these strokes other than the radical '门', they

are crowded together in tiny sizes and contain subtle differences, which could make the reconstruction task even more challenging.

5.4. Experiments for Zero-Shot Chinese Character Recognition

We employ the architecture based on ResNet and Transformer pre-trained on the second form of datasets for fine-tuning, as introduced in Section 4.2. The pre-training model for different zero-shot tasks should be trained on the corresponding class of datasets, i.e., the first 500, 1000, 1500, 2000 and 2755 classes of 3755 characters.

We compare our results with DenseRAN [12], HDE [13], and stroke-based methods [1]. The recognition accuracy of our method and other existing methods is shown in Table 1. The results show that our method outperforms other methods on various kinds of data sets. Our pre-training method achieves greater improvements on larger datasets compared to stroke-based method. What is more encouraging is, our SAE is pre-trained on a small and concise dataset with printed Chinese character images, but boosts huge performance in zero-shot recognition on a large and complicated dataset consisting of handwritten Chinese character images. To be specific, our SAE learns the stroke sequence of Chinese characters and spatial information of Chinese characters through masked image encoding, as a result, our SAE leverages the model to learn a better representation of handwritten Chinese characters at a higher semantic level.

Table 1. The results of zero-shot recognition with various approaches. Our model utilize the SAE based on ResNet and Transformer pre-trained on the second form of data sets.

Handwritten	**CCR Accuracy under Different Seen Classes**					**Printed Artistic**	**CCR Accuracy under Different Seen Classes**				
	500	**1000**	**1500**	**2000**	**2755**		**500**	**1000**	**1500**	**2000**	**2755**
DenseRAN [12]	1.70%	8.44%	14.71%	19.51%	30.68%	DenseRAN [12]	0.20%	2.26%	7.89%	10.86%	24.80%
HDE [13]	4.90%	12.77%	19.25%	25.13%	33.49%	HDE [13]	7.48%	21.13%	31.75%	40.43%	51.41%
Stroke-based [1]	5.60%	13.85%	22.88%	25.73%	37.91%	Stroke-based [1]	7.03%	26.22%	48.42%	54.86%	65.44%
Ours	**5.91%**	**14.35%**	**24.32%**	**30.17%**	**40.22%**	Ours	**8.25%**	**32.24%**	**50.72%**	**57.13%**	**68.88%**

5.5. Experiments for Seen Chinese Character Recognition

We pre-train our model on 3755 classes of Chinese characters. Then, the model is trained on HWDB1.0-1.1 and tested on full sets of ICDAR2013. At last, we combine the training parameters and pre-training parameters to fine-tune the model. The result is illustrated as SAE in the third block of Table 2. The first block of Table 2 shows the performance of human-beings [41] and character-based methods [44–47] and the second block presents some few/zero-shot methods [1,2,12,13,17,48]. The experimental results show that our SAE still outperforms few/zero-shot methods and some character-based methods, although it does not surpass all the character-based methods. This is probably because our model is mainly focusing on zero-shot recognition and therefore fails to achieve ideal results on seen recognition task.

Table 2. The results of seen Chinese character recognition with various approaches.

Method	**Accuracy (%)**
Human Performance [41]	96.13
HCCR-GoogLeNet [44]	96.35
DirectMap+ConvNet+Adaption [45]	97.37
M-RBC+IR [46]	97.37
Template+Instance [47]	97.45
DenseRAN [12]	96.66
FewShotRAN [48]	96.97
HDE [13]	96.74
Stroke-based [1]	96.28
REZCR [17]	96.35
STAR [2]	97.11
Ours	97.19

6. Discussion and Limitations

Intrinsic evaluation for word embeddings. Our pre-trained SAE model is competent in word embeddings. We use the architecture based on Vision Transformer pre-trained on 3755 printed Chinese character sequences to output word embeddings. We notice that the output of the encoder is 25 vectors in 64 dimension, and each word vector corresponds to a spatial position of the image. A 1600 (25×64) vector is too large for semantic embedding. For Chinese characters that have the same radical part at different locations, the embeddings of words cannot clearly show their relationship. In addition, each output of the decoder corresponds to an image of Chinese character's stroke sequence, and therefore the middle layer of the decoder contains the stroke information of Chinese characters. For a Chinese character and its stroke number m, the m-th output corresponds exactly to its original image.

Based on the above analysis, we use the m-th output patch of the third Transformer block of the decoder as its embedding vector, since it can reflect not only the stroke variation but also the overall structure of Chinese characters. To validate the representation ability of word embeddings of Chinese character structure, we calculate the word similarity of Chinese characters and their radical parts as an intrinsic evaluation. As shown in Figure 12, the Chinese characters and their corresponding radical parts have a higher cosine similarity in that row and column. This illustrates that our encoder enhances the representation of Chinese characters with their abundant morphological information.

	口	艹	辶	鸟	冬	加	召
鸣	**0.371**	0.477	0.442	**0.555**	0.441	0.444	0.349
图	**0.246**	0.371	0.348	0.379	**0.433**	0.351	0.285
茄	0.414	**0.493**	0.374	0.491	0.444	**0.583**	0.366
迢	0.429	0.357	**0.643**	0.327	0.387	0.262	**0.471**
莺	0.171	**0.510**	0.307	**0.621**	0.392	0.342	0.363
咚	**0.465**	0.398	0.461	0.522	**0.443**	0.346	0.380
迦	0.321	0.442	**0.582**	0.447	0.407	**0.451**	0.381
苕	0.453	**0.595**	0.303	0.355	0.529	0.420	**0.398**

Figure 12. The similarity between the Chinese characters and their radical parts. Bold numbers indicate the cosine similarity between Chinese characters and their corresponding radical components, and those enclosed in red boxes represent some outliers.

Moreover, we notice that the bold numbers in the first column do not reflect the strong relevance of Chinese characters and their radical parts. This is because all Chinese characters in the examples are composed of the same radical component '口', except '莺'. It can be seen that '莺' has the lowest cosine similarity in the first column, which is consistent with the above observation.

We use the t-SNE visualization [49] for the 256-dim embeddings of some Chinese characters. The K-means clustering algorithm [50] with cosine distance is applied to divide these characters into three categories, and each category is marked with the same color. As shown in Figure 13, we can see that these Chinese characters in the same category really have some common radical and stroke parts. Thus, the clustering results show great potential for our word embedding ability in other Chinese NLP-related applications, such as Chinese Text Classification [30] and Chinese Named Entities [6]. Our discussion of word embeddings from SAE architecture only scratches the surface of the Chinese character structure in semantic enhancement. How to replace the image representation

method in [30] or the Five-Stroke method in [6] with SAE is beyond the scope of this work, but worthy of further exploration.

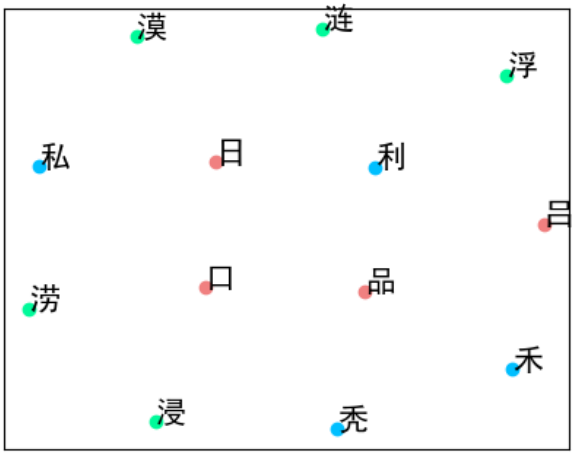


Figure 13. T-SNE visualization of embeddings learned from SAE architecture (different colors represent different classes).

Complexity analysis. We analysis the model size of two different models on training, pre-training and fine-tuning. For simplicity, we denote SAE based on Vision Transformer by SAE-ViT and SAE based on Resnet and Transformer by SAE-RT. The two training models are as large as stroke-based [1] and MAE [21] respectively. The details are shown in Table 3.

Table 3. The model sizes of our two different models on training, pre-training and fine-tuning.

Models	Train	Pre-Train	Fine-Tune
SAE-ViT	41.3 MB	43.6 MB	
SAE-RT	677 MB	679 MB	355 MB

Limitations. Our model successfully captures the morphological and semantic information of Chinese characters with self-supervised learning. However, since our model implements the self-supervised method by reconstructing stroke image sequences, how to assemble a high-quality dataset still remains challenging when we migrate the model to ancient Chinese characters or other hieroglyphs. That is, we need the stroke image sequences of the corresponding characters to implement the model training. Moreover, as a pre-training method, our model improves the performance based on stroke-level methods, which also means it relies on the performance of existing zero-shot recognition methods to a great extent.

7. Conclusions

In this paper, we propose stroke-based autoencoders to learn the structural information of Chinese characters. we propose two contrasting SAE architectures on different forms of data sets, which are employed for the unseen (zero-shot) and seen handwritten Chinese character recognition task and semantic enhancement task. Through combining our pre-training method and existing method, the proposed approach manages to achieve improved recognition accuracy results over existing methods and boost impressive performance on a large and complicated dataset by pre-training on a small and concise dataset. The experimental results on the benchmark datasets compared with existing methods validate that our SAE can effectively deal with the unseen (zero-shot) and seen Chinese character recognition problems. Among future directions, one particularly interesting extension is

to facilitate some related Chinese NLP tasks or migrate our model to the recognition of other hieroglyphs.

Author Contributions: Conceptualization, Z.C. and W.Y.; methodology, Z.C., W.Y. and X.L.; software, Z.C.; validation, Z.C., W.Y. and X.L.; formal analysis, Z.C., W.Y. and X.L.; investigation, Z.C., W.Y. and X.L.; writing—original draft preparation, Z.C.; writing—review and editing, Z.C., W.Y. and X.L.; visualization, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Natural Science Foundation of China grant number 61573012, China Scholarship Council grant number 201906955038, and National innovation and entrepreneurship training program for college students grant number 202210497065. The APC was funded by National innovation and entrepreneurship training program for college students grant number 202210497065.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some data presented in this study are openly available in [1,40,41]. And publicly available datasets were analyzed in this study. This data can be found here: [https://hanziwriter.org/].

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

References

1. Chen, J.; Li, B.; Xue, X. Zero-Shot Chinese Character Recognition with Stroke-Level Decomposition. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, Virtual, 19–26 August 2021.
2. Zeng, J.; Xu, R.; Wu, Y.; Li, H.; Lu, J. STAR: Zero-Shot Chinese Character Recognition with Stroke-and Radical-Level Decompositions. *arXiv* **2022**, arXiv:2210.08490.
3. Zeng, J.; Chen, Q.; Liu, Y.; Wang, M.; Yao, Y. StrokeGAN: Reducing Mode Collapse in Chinese Font Generation via Stroke Encoding. *Proc. AAAI Conf. Artif. Intell.* **2020**, *35*, 3270–3277. [CrossRef]
4. Wen, Q.; Li, S.; Han, B.; Yuan, Y. Zigan: Fine-grained chinese calligraphy font generation via a few-shot style transfer approach. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 621–629.
5. Wu, S.; Song, X.; Feng, Z. MECT: Multi-Metadata Embedding based Cross-Transformer for Chinese Named Entity Recognition. *arXiv* **2021**, arXiv:2107.05418.
6. Li, J.; Meng, K. MFE-NER: Multi-feature Fusion Embedding for Chinese Named Entity Recognition. *arXiv* **2021**, arXiv:abs/2109.07877.
7. Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; Li, J. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), Bangkok, Thailand, 1–6 August 2021.
8. Poddar, J.; Parikh, V.; Bharti, S.K. Offline signature recognition and forgery detection using deep learning. *Procedia Comput. Sci.* **2020**, *170*, 610–617. [CrossRef]
9. Ren, H.; Pan, M.; Li, Y.; Zhou, X.; Luo, J. ST-SiameseNet: Spatio-Temporal Siamese Networks for Human Mobility Signature Identification. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020.
10. Maekawa, K.; Tomiura, Y.; Fukuda, S.; Ishita, E.; Uchiyama, H. Improving OCR for historical documents by modeling image distortion. In Proceedings of the International Conference on Asian Digital Libraries, Kuala Lumpur, Malaysia, 4–7 November 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 312–316.
11. Huang, S.; Wang, H.; Liu, Y.; Shi, X.; Jin, L. Obc306: A large-scale oracle bone character recognition dataset. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 681–688.
12. Wang, W.; Zhang, J.; Du, J.; Wang, Z.R.; Zhu, Y. Denseran for offline handwritten chinese character recognition. In Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 104–109.
13. Cao, Z.; Lu, J.; Cui, S.; Zhang, C. Zero-shot Handwritten Chinese Character Recognition with hierarchical decomposition embedding. *Pattern Recognit.* **2020**, *107*, 107488. [CrossRef]
14. Zhang, J.; Zhu, Y.; Du, J.; Dai, L. Radical analysis network for zero-shot learning in printed Chinese character recognition. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.

15. Wang, S.; Huang, G.; Luo, X. Hippocampus-heuristic Character Recognition Network for Zero-shot Learning. *arXiv* **2021**, arXiv:2104.02236.
16. Huang, Y.; Jin, L.; Peng, D. Zero-Shot Chinese Text Recognition via Matching Class Embedding. In Proceedings of the Document Analysis and Recognition—ICDAR 2021, Lausanne, Switzerland, 5–10 September 2021; Lladós, J., Lopresti, D., Uchida, S., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 127–141.
17. Diao, X.; Shi, D.; Tang, H.; Li, Y.; Wu, L.; Xu, H. REZCR: Zero-shot Character Recognition via Radical Extraction. *arXiv* **2022**, arXiv:2207.05842.
18. Haralambous, Y. Seeking Meaning in a Space Made out of Strokes, Radicals, Characters and Compounds. *arXiv* **2011**, arXiv:1104.4321.
19. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition CVPR09, Miami, FL, USA, 20–25 June 2009.
20. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
21. He, K.; Chen, X.; Xie, S.; Li, Y.; Doll'ar, P.; Girshick, R.B. Masked Autoencoders Are Scalable Vision Learners. *arXiv* **2021**, arXiv:abs/2111.06377.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:abs/2010.11929.
23. Bao, H.; Dong, L.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. *arXiv* **2021**, arXiv:abs/2106.08254.
24. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
25. Chen, M.; Radford, A.; Wu, J.; Jun, H.; Dhariwal, P.; Luan, D.; Sutskever, I. Generative Pretraining From Pixels. In Proceedings of the Thirty-Seventh International Conference on Machine Learning ICML, Virtual Event, 13–18 July 2020.
26. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. *arXiv* **2021**, arXiv:abs/2102.12092.
27. Kodirov, E.; Xiang, T.; Gong, S. Semantic Autoencoder for Zero-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
28. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-Based Classification for Zero-Shot Visual Object Categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. [CrossRef]
29. Cao, S.; Lu, W.; Zhou, J.; Li, X. cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence AAAI, New Orleans, LA, USA, 2–7 February 2018.
30. Li, Y.; Zhao, Y.; Hu, B.; Chen, Q.; Xiang, Y.; Wang, X.; Ding, Y.; Ma, L. GlyphCRM: Bidirectional Encoder Representation for Chinese Character with its Glyph. *arXiv* **2021**, arXiv:abs/2107.00395.
31. Camacho-Collados, J.; Pilehvar, M.T.; Collier, N.; Navigli, R. *Semeval-2017 Task 2: Multilingual and Cross-Lingual Semantic Word Similarity*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017.
32. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical reasoning on chinese morphological and semantic relations. *arXiv* **2018**, arXiv:1805.06504.
33. Qiu, Y.; Li, H.; Li, S.; Jiang, Y.; Hu, R.; Yang, L. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 209–221.
34. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
35. Liu, P.D.; Chung, K.K.H.; McBride-Chang, C.A.; Tong, X. Holistic versus analytic processing: Evidence for a different approach to processing of Chinese at the word and character levels in Chinese children. *J. Exp. Child Psychol.* **2010**, *107*, 466–478. [CrossRef]
36. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:abs/1706.03762.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer Normalization. *arXiv* **2016**, arXiv:1607.06450.
39. Hanzi Writer. Available online: https://hanziwriter.org/ (accessed on 21 March 2022).
40. Liu, C.L.; Yin, F.; Wang, D.H.; Wang, Q.F. Online and offline handwritten Chinese character recognition: Benchmarking on new databases. *Pattern Recognit.* **2013**, *46*, 155–162. [CrossRef]
41. Yin, F.; Wang, Q.F.; Zhang, X.Y.; Liu, C.L. ICDAR 2013 Chinese Handwriting Recognition Competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1464–1470.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
43. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
44. Zhong, Z.; Jin, L.; Xie, Z. High performance offline handwritten chinese character recognition using googlenet and directional feature maps. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 846–850.

45. Zhang, X.Y.; Bengio, Y.; Liu, C.L. Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark. *Pattern Recognit.* **2017**, *61*, 348–360. [CrossRef]
46. Yang, X.; He, D.; Zhou, Z.; Kifer, D.; Giles, C.L. Improving offline handwritten Chinese character recognition by iterative refinement. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 1, pp. 5–10.
47. Xiao, Y.; Meng, D.; Lu, C.; Tang, C.K. Template-instance loss for offline handwritten Chinese character recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 315–322.
48. Wang, T.; Xie, Z.; Li, Z.; Jin, L.; Chen, X. Radical aggregation network for few-shot offline handwritten Chinese character recognition. *Pattern Recognit. Lett.* **2019**, *125*, 821–827. [CrossRef]
49. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
50. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.