Zero-shot Generation of Training Data with Denoising Diffusion Probabilistic Model for Handwritten Chinese Character Recognition

Dongnan Gui[1,2], Kai Chen[1(✉)], Haisong Ding[1], and Qiang Huo[1]

[1] Microsoft Research Asia, Beijing, China
{gdn2001,dinghs11}@mail.ustc.edu.cn, chenkai.cn@hotmail.com, qianghuo@microsoft.com
[2] University of Science and Technology of China, Hefei, China

Abstract. There are more than 80,000 character categories in Chinese while most of them are rarely used. To build a high performance handwritten Chinese character recognition (HCCR) system supporting the full character set with a traditional approach, many training samples need be collected for each character category, which is both time-consuming and expensive. In this paper, we propose a novel approach to transforming Chinese character glyph images generated from font libraries to handwritten ones with a denoising diffusion probabilistic model (DDPM). Training from handwritten samples of a small character set, the DDPM is capable of mapping printed strokes to handwritten ones, which makes it possible to generate photo-realistic and diverse style handwritten samples of unseen character categories. Combining DDPM-synthesized samples of unseen categories with real samples of other categories, we can build an HCCR system to support the full character set. Experimental results on CASIA-HWDB dataset with 3,755 character categories show that the HCCR systems trained with synthetic samples perform similarly with the one trained with real samples in terms of recognition accuracy. The proposed method has the potential to address HCCR with a larger vocabulary.

Keywords: Denoising Diffusion Probabilistic Model · Handwritten Chinese Character Recognition · Zero-shot Generation

1 Introduction

In the latest National Standards of the People's Republic of China about Chinese coded character set (GB18030-2022), 87,887 Chinese character categories are included. To create a high-performance handwritten Chinese character recognition (HCCR) system that supports the full character set using traditional

K. Chen—This work was done when Dongnan Gui was an intern in MMI Group, Microsoft Research Asia, Beijing, China.

G. A. Fink et al. (Eds.): ICDAR 2023, LNCS 14188, pp. 348–365, 2023.
https://doi.org/10.1007/978-3-031-41679-8_20

approaches, a large number of training samples with various writing styles would be collected for each character category. However, only about 4,000 categories are commonly used in daily life. It is therefore both time-consuming and expensive to collect representative handwritten samples for the remaining 95% rarely-used ones. These categories are often of complicated structures, existing in personal names, addresses, ancient books, historic documents and scientific publications. An HCCR system supporting the full-set of these categories with high accuracy will be beneficial to improve user experience, protect cultural heritages and promote academic exchanges.

Lots of research efforts have been made to build an HCCR system with only real training samples from commonly used characters. A Chinese character consists of radicals/strokes with specific spatial relationships, which are shared across all characters. Rather than encoding each character category as a single one-hot vector, [4,10,44,45] encode it as a sequence of radicals/strokes and spatial relationships to achieve zero-shot recognition goal. In [1,19,21,22], font-rendered glyph images are leveraged to provide reference representations for unseen character categories. There are also some efforts to synthesize handwritten samples for unseen categories. For example, [48] synthesizes unseen character samples with a radical composition network and combines them with real samples to train an HCCR system. However, its recognition accuracy is relatively poor.

We propose to solve this problem by synthesizing diverse and high-quality training samples for unseen character categories with denoising diffusion probabilistic models (DDPMs) [15,38]. Diffusion models have been shown to outperform other generation techniques in terms of diversity and quality [9,29,40–42], due to their powerful modeling capacity of high-dimensional distributions. This also offers a zero-shot generation capability. For example, in diffusion-based text-to-image generation [28,33,36], with all object types and spatial relationships existed in training samples, diffusion models are capable of generating photo-realistic images of in-existence object combinations and layouts. As mentioned above, Chinese characters can be treated as combinations of different radicals/strokes with specific layouts. We can leverage DDPM to achieve the goal of zero-shot handwritten Chinese character image generation.

In this paper, we design a glyph conditional DDPM (GC-DDPM), which concatenates a font-rendered character glyph image with the original input of U-Net used in [9], to guide the model in constructing mappings between font-rendered and handwritten strokes/radicals. To the best of our knowledge, we are the first to apply DDPMs to zero-shot handwritten Chinese character generation. Unlike other image-to-image diffusion model frameworks (e.g., [30,35,43]), which aim at synthesizing images in the target domain while faithfully preserving the content representations, our goal is to learn mappings from rendered printed radicals/strokes to the handwritten ones.

Experimental results on CASIA-HWDB [23] dataset with 3,755 character categories show that the HCCR systems trained with DDPM-synthesized samples outperform other synthetic data based solutions and perform similarly with the one trained with real samples in terms of recognition accuracy. We also

visualize the generation effect of both in and out of 3,755 character categories, which indicates that our method has the potential to be extended to a larger vocabulary.

The remainder of the paper is organized as follows. In Sect. 2, we briefly review related works. In Sect. 3, we describe our GC-DDPM design along with sampling methods. Our approach is evaluated and compared with prior arts in Sect. 4. We discuss limitations of our approach and future work in Sect. 5, and conclude the paper in Sect. 6.

2 Related Work

Zero-shot HCCR. Conventional HCCR systems [6,7,20,50,52,53], although achieving superior recognition accuracy, can only recognize character categories that are observed in the training set. Zero-shot HCCR aims to recognize handwritten characters that are never observed. Most of the previous zero-shot HCCR systems can be divided into two categories: structure-based and structure-free methods. In structure-based methods, a Chinese character is represented as a sequence of composing radicals [4,10,44,45] or strokes [5]. Although the character is never observed, the composing radicals, strokes and their spatial relationships have been observed in the training set. Therefore, structure-based methods are able to predict the radical or stroke sequences of unseen Chinese characters and achieve zero-shot recognition. However, in these methods, the radical or stroke sequence representations of Chinese characters require lots of language-specific domain knowledge. In structure-free method, [1,17,21,22] leverage information from the corresponding Chinese character glyph images. Zero-shot HCCR is achieved by choosing the Chinese character whose glyph features are closest to that of the handwritten ones in terms of visual representations. In [19], the radical information is also used to extract the visual representations of glyph images.

Zero-shot Data Synthesis for HCCR. Besides designing zero-shot recognition systems, there are some studies to directly synthesize handwritten training samples for unseen categories. [48] investigates a radical composition network to generate unseen Chinese characters by integrating radicals and their spatial relationships. Although the generated handwritten Chinese characters can increase the recognition rate of unseen handwritten characters, the overall recognition performance is relatively poor. In this work, we propose to use a more powerful diffusion model to generate unseen handwritten Chinese characters given corresponding glyph images.

Zero-shot Chinese Font Generation. Zero-shot Chinese font generation aims to generate font glyph for unseen Chinese characters based on some seen character/font glyph pairs. In [11,25,47,51,54], the image-to-image translation framework is used to achieve this goal. Works in [18,24,31] also leverage the information of composing components, radicals, strokes for better generalization. In this paper, we focus on zero-shot handwritten Chinese character generation with DDPM and we can easily adapt this method to zero-shot Chinese font generation task.

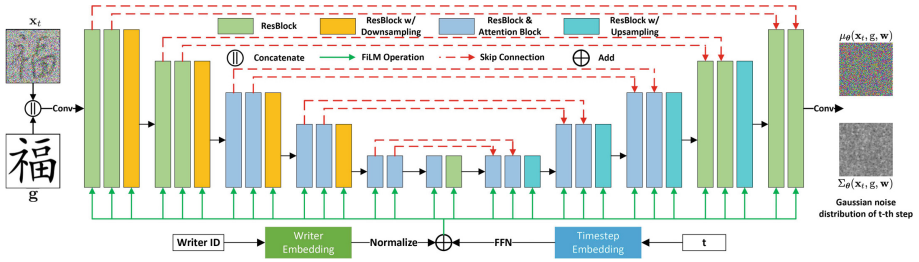


Fig. 1. Architecture of glyph conditional U-Net, which is adapted from the model used in [9]. We concatenate font "kai" rendered character image with original input to provide glyph guidance during generation.

Diffusion Model. DDPM [15,38] has become extremely popular in computer vision and achieves superior performance in image generation tasks. DDPM uses two parameterized Markov chains and variational inference method to reconstruct the data distribution. DDPMs have demonstrated their powerful capabilities to generate high-quality and high-diversity images [9,15,42]. It is shown in [33] that DDPM can perform a great effect on combination of concepts, which can integrate multiple elements. Diffusion models are also applied to other tasks [8,49], including high-resolution generation [34], image inpainting [43], natural language processing [2] and so on. Besides, [27] introduces DDPM to solve the problem of online English handwriting generation. In this work, we propose to leverage DDPM for zero-shot handwritten Chinese character generation and to synthesize training data for unseen Chinese characters to build HCCR systems.

3 Our Approach

3.1 Preliminary

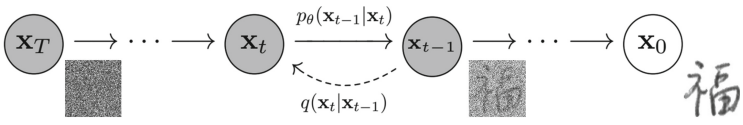


Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a handwritten Chinese character sample by slowly adding (removing) noise. Adapted from [15].

Diffusion model is a new paradigm of data generation. It defines a Markov chain of diffusion steps to slowly add random noise to data and then learn to reverse the diffusion process to construct desired data samples from the noise [46]. As

shown in Fig. 2, in our handwritten Chinese character generation scenario, we first sample a character image from the real distribution $\mathbf{x}_0 \sim q(\mathbf{x})$. Then, in forward diffusion process, small amounts of Gaussian noise are added to the sample in steps according to Eq. (1),

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \\ \mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t \end{aligned} \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, producing a sequence of noisy samples. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. As t becomes larger, the image gradually loses its distinguishable features. When $t \to \infty$, $\mathbf{x}_t$ becomes a sample of an isotropic Gaussian distribution.

If we can reverse the above process and sample from $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, we will be able to recreate the true sample from a Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If β_t is small enough, $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ will also be a Gaussian. So we can approximate it with a parameterized model, as shown in Eq. (2)

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) \ . \tag{2}$$

Since $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ is tractable,

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and

$$\tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t) \tag{4}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \ . \tag{5}$$

So we can train a neural network to approximate $\boldsymbol{\epsilon}_t$ and the predicted value is denoted as $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t)$. It has been verified that instead of directly setting $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ as $\tilde{\beta}_t$, setting it as a learnable interpolation between $\tilde{\beta}_t$, β_t in log domain will yield better log-likelihood [29]:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = \exp(\boldsymbol{\nu}_\theta(\mathbf{x}_t) \log \beta_t + (1 - \boldsymbol{\nu}_\theta(\mathbf{x}_t)) \log \tilde{\beta}_t) \ . \tag{6}$$

In this paper, we will train a U-Net to predict $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t)$ and $\boldsymbol{\nu}_\theta(\mathbf{x}_t)$ with the same hybrid loss as in [29].

3.2 Glyph Conditional U-Net Architecture

As shown in Fig. 1, the U-Net architecture we used is borrowed from [9]. With 128×128 image input, there are 5 resolution stages in encoder and decoder respectively, and each stage consists of 2 BigGAN residual blocks (ResBlock) [3]. In addition, BigGAN ResBlocks are also used for downsampling and upsampling activations. We also follow [9] to use multi-head attention at 32×32, 16×16

and 8×8 resolutions. Timestep t will first be mapped to sinusoidal embedding and then processed by a 2-layer feed-forward network (FFN). This processed embedding will then be fed to each convolution layer in U-Net through a feature-wise linear modulation (FiLM) operator [32].

To control the style and content of generated character images, writer information [12] and character category information are also fed to the model. Given a writer $\mathbf{w}$, which is actually the class index of all writer IDs, it will be mapped to a learnable embedding, followed by L2-normalization (denoted as $\mathbf{z}$), which is injected to U-Net together with the timestep embedding [29] as shown in Fig. 1.

If we inject character category information in the same way as writer, the model will not be able to generate samples for unseen categories because their embeddings are not optimized at all. In this paper, we propose to leverage printed images rendered by font "kai" to provide character category information. We denote this glyph image as $\mathbf{g}$. There are several ways to inject $\mathbf{g}$ to the model. For example, it can be encoded as a feature vector by a CNN/ViT and fed to U-Net in FiLM way, or encoded as feature sequences and fed to attention layers of U-Net serving as external keys and values [28]. In this paper, we simply inject $\mathbf{g}$ as model's input by concatenating it with $\mathbf{x}_t$ and leave other ways as future work. We call our approach as **G**lyph **C**onditional DDPM (**GC-DDPM**).

By conditioning model output on glyph image, we expect the model can learn the implicit mapping rules between printed stroke combinations and their handwritten counterparts. Then we can input font-rendered glyph images of unseen characters to the well-trained GC-DDPM and get their handwritten samples of high quality and diversity.

3.3 Multi-conditional Classifier-free Diffusion Guidance

Classifier-free guidance [16] has been proven effective for improving generation quality on different tasks. In this paper, we are also curious about its effects on HCCR system trained with synthetic samples.

There are 2 conditions, glyph $\mathbf{g}$ and writer $\mathbf{w}$, in our model. We assume that given $\mathbf{x}_t$, $\mathbf{g}$ and $\mathbf{w}$ are independent. So we have

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{g}, \mathbf{w}) \propto p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t) p_{\boldsymbol{\theta}}(\mathbf{g}|\mathbf{x}_t) p_{\boldsymbol{\theta}}(\mathbf{w}|\mathbf{x}_t) \ . \tag{7}$$

Following the previous practice in [16], we assume that there is an implicit classifier (ic),

$$p_{ic}(\mathbf{g}, \mathbf{w}|\mathbf{x}_t) \propto \left[\frac{p(\mathbf{x}_t|\mathbf{g})}{p(\mathbf{x}_t)}\right]^{\gamma} \cdot \left[\frac{p(\mathbf{x}_t|\mathbf{w})}{p(\mathbf{x}_t)}\right]^{\eta} \ . \tag{8}$$

Then we have

$$\nabla_{\mathbf{x}_t} \log p_{ic}(\mathbf{g}, \mathbf{w}|\mathbf{x}_t) \propto \gamma \boldsymbol{\epsilon}(\mathbf{x}_t, \mathbf{g}) + \eta \boldsymbol{\epsilon}(\mathbf{x}_t, \mathbf{w}) - (\gamma + \eta) \boldsymbol{\epsilon}(\mathbf{x}_t) \ . \tag{9}$$

So we can perform sampling with the score formulation

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{g}, \mathbf{w}) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{g}, \mathbf{w}) + \gamma \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{g}, \emptyset) \\ + \eta \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \emptyset, \mathbf{w}) - (\gamma + \eta) \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \emptyset, \emptyset) \ . \end{aligned} \tag{10}$$

We call γ, η as content and writer guidance scales respectively. When $\mathbf{g} = \emptyset$, an empty glyph image will be fed to U-Net and when $\mathbf{w} = \emptyset$, a special embedding will be used. During training, we set $\mathbf{g}$ and $\mathbf{w}$ to $\emptyset$ with probability 10% independently to get partial/unconditional models.

3.4 Writer Interpolation

Besides generating unseen characters, our model is also able to generate unseen styles by injecting interpolation between different writer embeddings as new writer embedding. Given two normalized writer embeddings $\mathbf{z_i}$ and $\mathbf{z_j}$, we use spherical interpolation [33] to get a new embedding $\mathbf{z}$ with L2-norm being 1, as in Eq. 11:

$$\mathbf{z} = \mathbf{z}_i \cos\frac{\lambda\pi}{2} + \mathbf{z}_j \sin\frac{\lambda\pi}{2}, \quad \lambda \in [0, 1]\ . \tag{11}$$

4 Experiments

We conduct our experiments on CASIA-HWDB [23] dataset. The detailed experimental setup is comprehensively explained in Sect. 4.1. Experiments on Writer Independent (WI) and Writer Dependent (WD) GC-DDPMs are conducted in Sect. 4.2 and Sect. 4.3, respectively. We further use synthesized samples to augment the training set of HCCR in Sect. 4.4. Finally, we compare our approach with prior arts in Sect. 4.5.

4.1 Experimental Setup

Dataset: The CASIA-HWDB dataset is a large-scale offline Chinese handwritten character database including HWDB1.0, 1.1 and 1.2. We use the HWDB1.0 and 1.1 in experiments, where the former contains 3,866 Chinese character categories written by 420 writers, and the latter contains 3,755 categories written by another 300 writers. We follow the official partition of training and testing sets as in [23], where the training set is written by 576 writers.

Vocabulary Partition: We use the 3,755 categories that cover the standard GB2312-80 level-1 Chinese set in experiments. We denote the set of 3,755 categories as $\mathcal{S}_{3,755}$. Following the setup in [1,45], we select the first 2,000 categories in GB2312-80 set as seen categories (denoted as $\mathcal{S}_{2,000}$), and the remaining 1,755 categories as unseen categories (denoted as $\mathcal{S}_{1,755}$). The diffusion models are trained on training samples of $\mathcal{S}_{2,000}$ and used to generate handwritten Chinese character samples of $\mathcal{S}_{1,755}$ to evaluate the performance of zero-shot training data generation for HCCR.

DDPM Settings: Our DDPM implementation is based on [9]. We use the "kai" as our font library to render printed character images. We conduct experiments on both WI and WD GC-DDPMs. In WI GC-DDPM training, we disable writer embeddings and randomly set content condition $\mathbf{g}$ as $\emptyset$ with probability 10%.

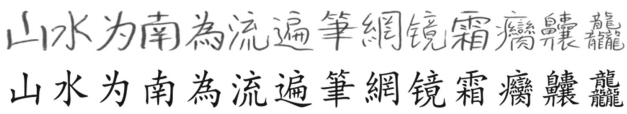


Fig. 3. Synthetic handwritten Chinese character samples and corresponding glyphs, with stroke numbers increasing from left to right.

And in WD GC-DDPM, writer condition $\mathbf{w}$ is also randomly set to $\emptyset$ with probability 10%. Flip and mirror augmentations are used during training. We set batch size as 256, image size as 128×128, and we use AdamW optimizer [26] with learning rate 1.0e-4. Diffusion step number is set to 1,000 with a linear noise schedule. GC-DDPMs are trained for about 200K steps using a machine with 8 Nvidia V100 GPUs, which takes about 5 d. During sampling, we use the denoising diffusion implicit model (DDIM) [39] sampling method with 50 steps. It takes 62 h to sample 3,755 characters written by 576 writers, which are about 2.2M samples, with the same 8 Nvidia V100 GPUs.

Evaluation Metrics: We evaluate the quality of synthetic samples in three aspects. First, Inception score (IS) [37] and Frechet Inception Distance (FID) [14] are used to evaluate the diversity and distribution similarity of synthetic samples compared with real ones. Second, since samples are synthesized by conditioning on glyph image, the synthetic samples should be consistent with the category of conditioned glyph. Therefore, we introduce a new metric called correctness score (CS). For each synthetic sample, the category of conditioned glyph is used as ground truth, and CS is calculated as the recognition accuracy of synthetic samples using an HCCR model trained with real data, which achieves 97.3% recognition accuracy in real data testing set. Finally, as the purpose of diffusion model here is to generate training data for unseen categories, we also train HCCR models with synthetic samples and evaluate recognition accuracy on the real testing set of unseen categories. Our HCCR model adopts ResNet-18 [13] architecture and is trained with standard SGD optimizer. No data augmentation is applied during HCCR model training. It is noted that starting from different random noise, it is almost impossible to generate exact same handwritten samples even for same conditional character glyphs. So it is not appropriate to adopt pixel-level metrics to evaluate generative effect as [11,18,24,25,31,47,51,54] do (Fig. 3).

4.2 WI GC-DDPM Results

We first conduct experiments on WI GC-DDPM. It is shown in [16] that the classifier guidance scale is able to attain a trade-off between quality and diversity. In order to evaluate the behavior of different content guidance scale γ's, we choose different γ's and generate samples to compute FID, ID and CS. Here we synthesize 50K samples of $\mathcal{S}_{2,000}$, and the HCCR model used to measure CS is trained using real samples of $\mathcal{S}_{3,755}$. $\gamma \in \{0.0,\ 1.0,\ 2.0,\ 3.0\ , 4.0\}$ are used and the comparison results are summarized in Table 1. We can find that, as γ

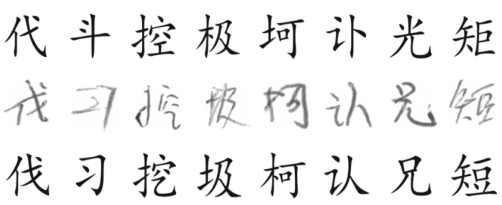
Table 1. Comparisons of generation quality using different content guidance scale γ's in terms of IS, FID, and CS.

γ	IS	FID	CS (%)
0.0	2.62	8.07	94.7
1.0	2.51	10.97	99.8
2.0	2.46	18.03	99.9
3.0	2.44	24.34	99.9
4.0	2.39	28.69	99.9

Table 2. Comparisons of generation quality using different content guidance scale γ's in terms of recognition accuracy on testing set of classes in $\mathcal{S}_{1,755}$ using generated samples as training set.

γ	0.0	1.0	2.0	3.0	4.0
$\text{Acc}_{1,755}$ (%)	93.0	88.6	91.7	63.7	33.2

(a) Failure samples that do not look like any Chinese characters.



(b) (top) Glyph condition images; (middle) Synthetic samples; (bottom) Most similar characters.

Fig. 4. Synthetic samples that are wrongly recognized by real data trained HCCR model when $\gamma = 0$.

increases, the IS decreases, the FID increases and the CS achieves close to 100% accuracy. This indicates that with a larger γ, the diversity of synthetic samples is decreasing. This behavior is also observed in Fig. 5a where we visualize multiple sampled results of the character class in $\mathcal{S}_{2,000}$ using different γ's. The generated samples are less diverse, less cursive and easier to recognize when conditioned on stronger content guidance. According to FID and examples in Fig. 5, the distribution of synthetic samples with $\gamma = 0$ is closer to that of real samples.

(a) $\mathcal{S}_{2,000}$ example

(b) $\mathcal{S}_{1,755}$ example

(c) Out of $\mathcal{S}_{3,755}$ example

(d) Complicated strokes example

Fig. 5. Multiple synthetic handwritten Chinese character samples with different content guidance scale, where (a), (b) and (c) are characters from classes of $\mathcal{S}_{2,000}$, $\mathcal{S}_{1,755}$, and out of $\mathcal{S}_{3,755}$ Chinese character sets. Samples in each line use the same random seed and initial noise. Samples across lines use different random seeds to visualize diversity.

When $\gamma = 0$, CS achieves 94.7%. In Fig. 4, we show synthetic cases that the trained HCCR model fails to recognize. Failure cases include (a) samples that are unreadable, and (b) samples that are closer to another easily confused Chinese character. They are caused by alignment failures between printed and synthetic strokes, and can be eliminated by improving glyph conditioning method. We leave it as future work.

Then, we evaluate the quality of WI GC-DDPM for zero-shot generation of HCCR training data. We use the trained WI GC-DDPM to synthesize 576 samples for each category in $\mathcal{S}_{1,755}$. Then, the synthetic samples are used along with real samples of categories in $\mathcal{S}_{2,000}$ to train an HCCR model that supports 3,755 categories. We calculate its recognition accuracy on the testing set of category $\mathcal{S}_{1,755}$, which is denoted as $\text{Acc}_{1,755}$. Different γ's are tried, and the results are shown in Table 2. In Fig. 5b, we visualize synthetic samples of one category in $\mathcal{S}_{1,755}$. The best $\text{Acc}_{1,755}$ is achieved when $\gamma = 0$. Although synthetic samples with higher γ are less cursive, they achieve much lower $\text{Acc}_{1,755}$. This is because the lack of diversity makes it difficult to cover the wide distribution of handwritten Chinese character image space.

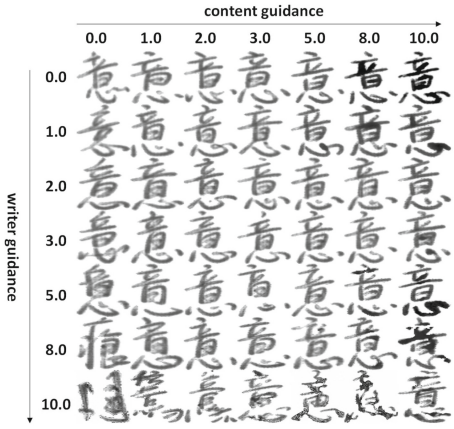


Fig. 6. Generated handwritten Chinese character samples with different content and writer guidance scales, where the character is from the class of $\mathcal{S}_{1,755}$. Samples are generated with the same random seed and initial noise.

Table 3. Comparisons of generation quality between WI and WD DDPMs in terms of IS, FID, CS (%) and the recognition accuracy (%) on the testing set of class $\mathcal{S}_{1,755}$ using generated samples as training set.

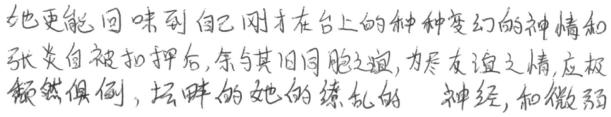
Model	IS	FID	CS	$\text{Acc}_{1,755}$
WI	2.62	8.07	94.7	93.0
WD	2.49	6.34	94.8	93.7
WD w/ interpolation	2.53	6.26	95.0	94.7

Clearly, by learning the mapping of radicals and spatial relationship between Chinese printed and handwritten strokes, the diffusion model is capable of zero-shot generation of unseen Chinese character categories. Moreover, a high accuracy of 93.0% is achieved on $\mathcal{S}_{1,755}$ by only leveraging the synthetic samples. In Figs. 5c and 5d, we further show the synthetic samples of a Chinese character category that does not belong to $\mathcal{S}_{3,755}$. The excellent generation effect implies that our method has the potential to be extended to a larger vocabulary.

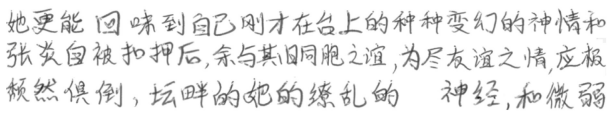
4.3 WD GC-DDPM Results

Although WI GC-DDPM can generate desired handwritten characters, we cannot control their writing styles. In this part, we conduct experiments on WD GC-DDPM, which introduces writer information as an additional condition.

Figure 6 shows the visualization results of sampling with different content guidance scale γ's and writer guidance scale η's. It shows that with larger γ, the synthetic samples become less cursive and more similar to the corresponding printed image. This behavior is consistent with that of the WI GC-DDPM in



(a) Real text line from [23].



(b) Synthetic samples arranged as a text line.

Fig. 7. Comparisons of real text line images in HWDB2.1 and generated samples arranged in a text line, where we replace the characters from real data with the generated characters. Samples in different lines of (a) and (b) are selected and generated conditioning on the same writer 1001.

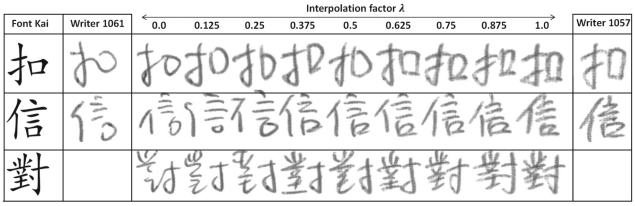


Fig. 8. Interpolation of handwritten Chinese character samples, where the top, middle, bottom lines are characters from classes of $\mathcal{S}_{2,000}$, $\mathcal{S}_{1,755}$, and out of $\mathcal{S}_{3,755}$ Chinese character sets. We choose writer 1061 (left) and writer 1057 (right) for interpolation and interpolation factors are shown at the top of images. Standard glyph images of font "kai" are shown on the left. Samples in each line use the same random seed and initial noise.

Fig. 5. We also find that with large η, the generated sample becomes inconsistent with the conditioned printed image. Since writer information is injected to GC-DDPM in FiLM way, a large guidance scale will cause the mean and variance shift of $\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{g}, \mathbf{w})$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{g}, \mathbf{w})$ which hinders the subsequent denoising, leading to over-saturated images with over-smoothed textures [43].

In Fig. 7b, we show several synthetic text line images conditioned on a fixed writer embedding with our WD GC-DDPM. Writing styles of these samples are consistent and quite similar to real samples written by the same writer as shown in Fig. 7a. These results verify the writing style controllability of our model.

Then, we compare the quality of synthetic samples when used as training data for HCCR. For a fair comparison, we also generate 576 samples for each category in $\mathcal{S}_{1,755}$, one image for each writer. Recognition performances are shown in Table 3. To improve sampling efficiency and ensure training data diversity, the writer guidance scale of 0 is applied. Compared with using samples synthesized with WI GC-DDPM as HCCR training set, the accuracy on the testing set

Table 4. Comparisons of recognition accuracy (%) on test sets of $\mathcal{S}_{2,000}$ and $\mathcal{S}_{1,755}$ using real and/or synthetic samples as HCCR training set.

Training set		Accuracy on testing set	
Real	Synthetic	$\text{Acc}_{2,000}$	$\text{Acc}_{1,755}$
✓	/	97.3	97.2
/	WI	96.3	96.0
/	WD	96.4	96.1
/	WD w/ interpolation	96.5	96.1
✓	WI	97.3	97.3
✓	WD	97.4	97.3
✓	WD w/ interpolation	97.4	97.3

of $\mathcal{S}_{1,755}$ is improved from 93.0% to 93.7%. When GC-DDPM is trained without conditioning on writer embedding, it may generate similar samples from different initial noise. Whereas in WD GC-DDPM, by conditioning on different writer embeddings, the model will generate samples with different writing styles. Therefore, the diversity of synthetic samples will be improved. To verify this, we compare the quality of synthetic samples in terms of IS and FID. As shown in Table 3, the FID improves from 8.07 to 6.34. The results demonstrate the superiority of WD GC-DDPM in zero-shot training data generation of unseen Chinese character categories.

Another capability of WD GC-DDPM is that it can interpolate between different writer embeddings and generate samples of new styles. We choose 2 writers and try different interpolation factor λ's and visualize the synthetic samples in Fig. 8. We find that as λ increases from 0 to 1, the style of synthetic samples gradually shifts from one writing style to another. We also observe that with the same λ, the synthetic samples of different Chinese characters share similar writing style as expected. Finally, we use writer style interpolation to generate the training data of $\mathcal{S}_{1,755}$ for HCCR, and again 576 samples are generated for each category. For each image, we randomly select 2 writers for interpolation. We simply use an interpolation factor of 0.5. Results are summarized in Table 3. We observe a slight improvement in FID score and a 1% absolute recognition accuracy improvement on $\mathcal{S}_{1,755}$, which further verifies the superiority of our WD GC-DDPM.

4.4 Data-Augmented HCCR Results

We also use GC-DDPMs trained on $\mathcal{S}_{2,000}$, to synthesize samples for all categories in $\mathcal{S}_{3,755}$, and combine them with real samples to build HCCR systems. 3 settings are tried: WI, WD and WD w/ interpolation. And 576 samples for each category are synthesized in each setting. Table 4 summarizes the results. Best accuracies are achieved with samples synthesized by WD w/ interpolation,

Table 5. Comparisons of unseen character categories' recognition accuracy (%) between our method and prior zero-shot HCCR systems. Works with * also use samples from HWDB1.2 for training, while † means online trajectory information is also used.

Method	Accuracy
CM† [1]	86.7
DenseRan [45]	19.5
FewRan* [44]	70.6
HCCR* [4]	73.4
OSOCR* [21]	84.3
OSCCD* [22]	95.6
WI GC-DDPM	96.4
WD GC-DDPM	96.8
WD GC-DDPM w/interpolation	**96.9**

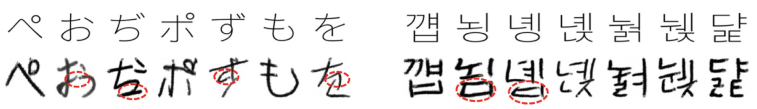
Table 6. Comparisons of unseen character categories' recognition accuracy (%) on CASIA1.2 testing set.

Methods	Accuracy
RCN [48]	46.1
WI GC-DDPM	98.6
WD GC-DDPM	98.6
ResNet-18 trained with real data	97.9

which is consistent with Table 3. The HCCR models trained with only synthetic samples perform slightly worse than the one trained with only real samples. Combining synthetic and real training samples only performs 0.0%˜0.1% better than real samples. These results demonstrate the distribution modeling capacity of GC-DDPMs.

4.5 Comparison with Prior Arts

Finally, we compare our method with prior arts. We first compare our method with prior zero-shot HCCR systems. To be consistent with prior works in [4,21,22], we randomly choose 1,000 classes in $\mathcal{S}_{1,755}$ as unseen classes and use ICDAR2013 [50] benchmark dataset for testing. Results are shown in Table 5. Here we only list the results from prior arts using 2,000 seen character classes. It is noted that the 2,000/1,000 seen/unseen character class split for training and testing is not exactly the same. So the results are not directly comparable. The results in Table 5 show that our methods achieve the same level recognition accuracy compared with previous state-of-the-art zero-shot HCCR systems. Moreover, our approach directly uses a standard CNN to predict supported categories, which is much simpler compared with the systems in [21,22].



(a) Japanese (b) Korean

Fig. 9. Synthetic samples of Japanese and Korean characters and standard glyph images in font "SourceHans".

We also compare our approach with [48], which also leverages a generation model to synthesize training samples for unseen classes. We follow the same experimental setups in [48] and use HWDB1.0 and 1.1 as training set, which contains 3,755 categories, to train GC-DDPMs. Unseen 3,319 categories in HWDB1.2 testing set are used as testing set. Results are shown in Table 6. [48] achieves a 46.1% accuracy by adding more than 9.6M generated samples. Our approach achieves a 98.6% accuracy by only adding about 1.9M synthetic samples (576 samples for each unseen category). We also train a classifier using all real samples in HWDB1.2 training set (240 samples for each category). The classifier achieves a 97.9% accuracy, which is slightly worse than ours due to less diverse training samples.

These results verify the zero-shot generation capability of our methods again. It is easy to extend to larger vocabularies, which makes it possible to build a high-quality HCCR system for 87,887 categories.

5 Limitations and Future Work

Although GC-DDPM-synthesized images are quite helpful for building a high-quality HCCR system, there are still some failure cases. The blur and dislocation phenomena in these samples reveal that there exist better ways to inject glyph information. It is also possible to encode radical/stroke sequences with spatial relationships as the condition of DDPM. We will investigate these methods and report the results elsewhere.

Another limitation of our approach is the long training time of DDPMs. We will try to reduce the number of character categories and sample numbers per category to find a better trade-off between synthesis quality and training cost.

Japanese and Korean characters share most strokes with Chinese, so we also try to synthesize handwritten Japanese and Korean samples with our Chinese-trained DDPM. As Fig. 9 shows, except for some circle and curve strokes, the results are quite reasonable. As future work, we will combine handwritten samples of CJK languages to build a new DDPM, which is expected to synthesize samples for each language with higher diversity and quality.

6 Conclusion

We propose WI and WD GC-DDPM solutions to achieve zero-shot training data generation for HCCR. Experimental results have verified their effectiveness in

terms of generation quality, diversity and HCCR accuracies of unseen categories. WD performs slightly better than WI due to its better distribution modeling capability and writing style controllability. These solutions can be easily extended to larger vocabularies and other languages, and provide a feasible way to build an HCCR system supporting 87,887 categories with high recognition accuracy.

References

1. Ao, X., Zhang, X.Y., Yang, H.M., Yin, F., Liu, C.L.: Cross-modal prototype learning for zero-shot handwriting recognition. In: ICDAR, pp. 589–594 (2019)
2. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., van den Berg, R.: Structured denoising diffusion models in discrete state-spaces. In: NeurIPS, vol. 34, pp. 17981–17993 (2021)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
4. Cao, Z., Lu, J., Cui, S., Zhang, C.: Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. Pattern Recogn. **107**, 107488 (2020)
5. Chen, J., Li, B., Xue, X.: Zero-shot Chinese character recognition with stroke-level decomposition. In: IJCAI, pp. 615–621 (2021)
6. Chen, L., Wang, S., Fan, W., Sun, J., Naoi, S.: Beyond human recognition: a CNN-based framework for handwritten character recognition. In: ACPR, pp. 695–699 (2015)
7. Cireşan, D., Meier, U.: Multi-column deep neural networks for offline handwritten Chinese character classification. In: IJCNN, pp. 1–6 (2015)
8. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: a survey. CoRR abs/2209.04747 (2022)
9. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS, vol. 34, pp. 8780–8794 (2021)
10. Diao, X., Shi, D., Tang, H., Wu, L., Li, Y., Xu, H.: REZCR: a zero-shot character recognition method via radical extraction. CoRR abs/2207.05842 (2022)
11. Gao, Y., Guo, Y., Lian, Z., Tang, Y., Xiao, J.: Artistic glyph image synthesis via one-stage few-shot learning. ACM TOG **38**(6), 1–12 (2019)
12. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS, vol. 30, pp. 6626–6637 (2017)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop DGMs Applications (2021)
17. Huang, G., Luo, X., Wang, S., Gu, T., Su, K.: Hippocampus-heuristic character recognition network for zero-shot learning in Chinese character recognition. Pattern Recogn. **130**, 108818 (2022)
18. Huang, Y., He, M., Jin, L., Wang, Y.: RD-GAN: few/zero-shot Chinese character style transfer via radical decomposition and rendering. In: ECCV, pp. 156–172 (2020)

19. Huang, Y., Jin, L., Peng, D.: Zero-shot Chinese text recognition via matching class embedding. In: ICDAR, pp. 127–141 (2021)
20. Li, Z., Teng, N., Jin, M., Lu, H.: Building efficient CNN architecture for offline handwritten Chinese character recognition. Int. J. Document Anal. Recog. **21**(4), 233–240 (2018)
21. Liu, C., Yang, C., Qin, H.B., Zhu, X., Liu, C.L., Yin, X.C.: Towards open-set text recognition via label-to-prototype learning. Pattern Recogn. **134**, 109109 (2022)
22. Liu, C., Yang, C., Yin, X.C.: Open-set text recognition via character-context decoupling. In: CVPR, pp. 4523–4532 (2022)
23. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: CASIA online and offline Chinese handwriting databases. In: ICDAR, pp. 37–41 (2011)
24. Liu, W., Liu, F., Ding, F., He, Q., Yi, Z.: XMP-Font: self-supervised cross-modality pre-training for few-shot font generation. In: CVPR, pp. 7905–7914 (2022)
25. Liu, Y., Lian, Z.: FontTransformer: few-shot high-resolution Chinese glyph image synthesis via stacked Transformers. CoRR abs/2210.06301 (2022)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
27. Luhman, T., Luhman, E.: Diffusion models for handwriting generation. CoRR abs/2011.06704 (2020)
28. Nichol, A., et al.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML, vol. 162, pp. 16784–16804 (2022)
29. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML, pp. 8162–8171 (2021)
30. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: methods and applications. IEEE Trans. Multimedia **24**, 3859–3881 (2021)
31. Park, S., Chun, S., Cha, J., Lee, B., Shim, H.: Few-shot font generation with localized style representations and factorization. In: AAAI, pp. 2393–2402 (2021)
32. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: visual reasoning with a general conditioning layer. In: AAAI, pp. 3942–3951 (2018)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. CoRR abs/2204.06125 (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)
35. Saharia, C., et al.: Palette: image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10 (2022)
36. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. CoRR abs/2205.11487 (2022)
37. Salimans, T., et al.: Improved techniques for training GANs. In: NeurIPS, vol. 29, pp. 2226–2234 (2016)
38. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML, pp. 2256–2265 (2015)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
40. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS, vol. 32, pp. 11895–11907 (2019)
41. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. In: NeurIPS, vol. 33, pp. 12438–12448 (2020)
42. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
43. Wang, T., et al.: Pretraining is all you need for image-to-image translation. CoRR abs/2205.12952 (2022)

44. Wang, T., Xie, Z., Li, Z., Jin, L., Chen, X.: Radical aggregation network for few-shot offline handwritten Chinese character recognition. Pattern Recogn. Lett. **125**, 821–827 (2019)
45. Wang, W., Zhang, J., Du, J., Wang, Z.R., Zhu, Y.: DenseRAN for offline handwritten Chinese character recognition. In: ICFHR, pp. 104–109 (2018)
46. Weng, L.: What are diffusion models? https://lilianweng.github.io/posts/2021-07-11-diffusion-models/, July 2021
47. Xie, Y., Chen, X., Sun, L., Lu, Y.: DG-Font: deformable generative networks for unsupervised font generation. In: CVPR, pp. 5130–5140 (2021)
48. Xue, M., Du, J., Zhang, J., Wang, Z.R., Wang, B., Ren, B.: Radical composition network for Chinese character generation. In: ICDAR, pp. 252–267 (2021)
49. Yang, L., et al.: Diffusion models: a comprehensive survey of methods and applications. CoRR abs/2209.00796 (2022)
50. Yin, F., Wang, Q.F., Zhang, X.Y., Liu, C.L.: ICDAR 2013 Chinese handwriting recognition competition. In: ICDAR, pp. 1464–1470 (2013)
51. Zhang, Y., Zhang, Y., Cai, W.: Separating style and content for generalized style transfer. In: CVPR, pp. 8447–8455 (2018)
52. Zhong, Z., Zhang, X.Y., Yin, F., Liu, C.L.: Handwritten Chinese character recognition with spatial Transformer and deep residual networks. In: ICPR, pp. 3440–3445 (2016)
53. Zhong, Z., Jin, L., Xie, Z.: High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps. In: ICDAR, pp. 846–850 (2015)
54. Zhu, A., Lu, X., Bai, X., Uchida, S., Iwana, B.K., Xiong, S.: Few-shot text style transfer via deep feature similarity. IEEE Trans. Image Process. **29**, 6932–6946 (2020)