FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning

Zhenhua Yang[1], Dezhi Peng[1], Yuxin Kong[1], Yuyi Zhang[1], Cong Yao[3], Lianwen Jin[12]*

[1]South China University of Technology
[2]SCUT-Zhuhai Institute of Modern Industrial Innovation
[3]Alibaba Group
{eezhyang, pengdezhi000, kongyxscut, yuyizhang.scut, yaocong2010}@gmail.com, eelwjin@scut.edu.cn

Abstract

Automatic font generation is an *imitation task*, which aims to create a font library that mimics the style of reference images while preserving the content from source images. Although existing font generation methods have achieved satisfactory performance, they still struggle with *complex characters* and *large style variations*. To address these issues, we propose **FontDiffuser**, a diffusion-based image-to-image one-shot font generation method, which innovatively models the font imitation task as a noise-to-denoise paradigm. In our method, we introduce a Multi-scale Content Aggregation (MCA) block, which effectively combines global and local content cues across different scales, leading to enhanced preservation of intricate strokes of complex characters. Moreover, to better manage the large variations in style transfer, we propose a Style Contrastive Refinement (SCR) module, which is a novel structure for style representation learning. It utilizes a style extractor to disentangle styles from images, subsequently supervising the diffusion model via a meticulously designed style contrastive loss. Extensive experiments demonstrate FontDiffuser's state-of-the-art performance in generating diverse characters and styles. It consistently excels on complex characters and large style changes compared to previous methods. The code is available at https://github.com/yeungchenwa/FontDiffuser.

Introduction

Automatic font generation aims to create a new font library in the required style given the reference images, which is referred to as an *imitation task*. Font generation has significant applications, including new font creation, ancient character restoration, and data augmentation for optical character recognition. Therefore, it has significant commercial and cultural values. However, this imitation process is both costly and labor-intensive, particularly for languages with a large number of glyphs, such as Chinese ($> 90,000$), Japanese ($> 50,000$), and Korean (> 11000). Existing automatic methods primarily disentangle the representations of style and content, then integrate them to output the results.

Although these methods have achieved remarkable success in font generation, they still suffer from *complex character generation* and *large style variation transfer*, leading

---
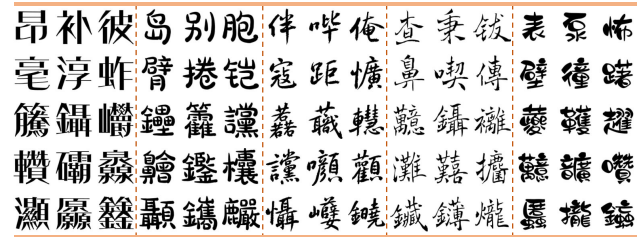
(a) Characters generated by our method

source ref (1) (2) (3) (4) **Ours**

(b) Complex characters

source ref (1) (2) (3) (4) **Ours**

(c) Large style variations

Figure 1: (a) Characters of different complexity generated by our method. (b)(c) Results of different methods on complex characters and large style variations. 'ref' represents the reference image. (1)-(4) represent the results of DG-Font (Xie et al. 2021), MX-Font (Park et al. 2021b), CG-GAN (Kong et al. 2022), and CF-Font (Wang et al. 2023) respectively. Red boxes highlight the failures of other methods.

to severe stroke missing, artifacts, blurriness, layout errors, and style inconsistency as shown in Figure 1(b)(c). Retrospectively, most font generation approaches (Park et al. 2021a,b; Xie et al. 2021; Tang et al. 2022; Liu et al. 2022; Kong et al. 2022; Wang et al. 2023) adopt a GAN-based (Goodfellow et al. 2014) framework which potentially suffers from unstable training due to their adversarial training nature. Moreover, most of these methods perceive content information through only single-scale high-level features, omitting the fine-grained details that are crucial to preserving the source content, especially for complex characters. There are also a number of methods (Cha et al. 2020; Park et al. 2021a,b; Liu et al. 2022; Kong et al. 2022; He et al. 2022) that employ prior knowledge to facilitate font generation, such as stroke or component composition of characters; however, this information is costly to annotate for complex characters. Furthermore, the target style is commonly represented by a simple classifier or a discriminator in previous literature, which struggles to learn the appropri-

ate style and hinders the style transfer with large variations.

In this paper, we propose FontDiffuser, a diffusion-based image-to-image one-shot font generation method, which models the font generation learning as a noise-to-denoise paradigm and is capable to generate unseen characters and styles. In our method, we innovatively introduce a Multi-scale Content Aggregation (MCA) block, which leverages global and local content features across various scales. This block effectively preserves intricate details from the source image of complex characters, by capitalizing on the fact that large-scale features contain lots of fine-grained information (strokes or components), whereas small-scale features primarily encapsulate global information (layout). Moreover, we introduce a novel style representation learning strategy, by applying a Style Contrastive Refinement (SCR) module to enhance the generator's capability in mimicking styles, especially for large variations between the source image and the reference image. This module utilizes a style extractor to disentangle style from a font and then uses a style contrastive loss to provide feedback to the diffusion model. SCR acts as a supervisor and encourages our diffusion model to identify the differences among various samples, which are with different styles but the same character. Additionally, we design a Reference-Structure Interaction (RSI) block to explicitly learn structural deformations (*e.g.*, font size) by utilizing a cross-attention interaction with the reference features.

To verify the effectiveness of generating characters of diverse complexity, we categorize the characters into three levels of complexity (easy, medium, and hard) according to their number of strokes, and test our method on each level separately. Extensive experiments demonstrate that our proposed FontDiffuser outperforms state-of-the-art font generation methods on characters of three levels of complexity. Notably, as shown in Figure 1(a), FontDiffuser consistently excels both in the generation of complex characters and large style variations. Furthermore, our method can be applied to the cross-lingual generation tasks, showcasing the cross-domain generalization ability of FontDiffuser.

We summarize our main contributions as follows.

- We propose FontDiffuser, a new diffusion-based image-to-image one-shot font generation framework that achieves state-of-the-art performance in generating complex characters and handling large style variations.

- To enhance the preservation of intricate strokes of complex characters, we propose a Multi-scale Content Aggregation (MCA) block, leveraging the global and local features across different scales from the content encoder.

- We propose a novel style representation learning strategy and elaborate a Style Contrastive Refinement (SCR) module that supervises the diffusion model using a style contrastive loss, enabling effective handling of large style variations.

- FontDiffuser demonstrates superior performance over existing methods in generating characters across easy, medium, and hard complexity levels, showcasing strong generalization capability across unseen characters and styles. Furthermore, our method can be extended to the cross-lingual generation, such as Chinese to Korean.

Related Work

Few-Shot Font Generation

Early font generation methods (Chang et al. 2018; Lyu et al. 2017; Tian 2017; Jiang et al. 2017; Sun, Zhang, and Yang 2018) consider the font generation task as an image-to-image translation problem, but they cannot generate unseen style fonts. To address this, SA-VAE (Sun et al. 2017) and EMD (Zhang, Zhang, and Cai 2018) generate unseen fonts by disentangling style and content representations. To enable the generator to capture local style characteristics, some methods (Wu, Yang, and Hsu 2020; Huang et al. 2020; Cha et al. 2020; Park et al. 2021a,b; Liu et al. 2022; Kong et al. 2022) utilize prior knowledge, such as stroke and component. For instance, LF-Font (Park et al. 2021a), MX-Font (Park et al. 2021b) and CG-GAN (Kong et al. 2022) employ a component-based learning strategy to enhance the capability of local style representation learning. XMP-Font (Liu et al. 2022) utilizes a pre-training strategy to facilitate the disentanglement of style and content. Diff-Font (He et al. 2022) adopts stroke information to support the sampling but fails to generate unseen characters. However, the annotation of strokes and components is costly for complex characters. Some prior-free methods (Xie et al. 2021; Tang et al. 2022; Wang et al. 2023) have been proposed. DG-Font (Xie et al. 2021) achieves promising performance in an unsupervised manner. Fs-Font (Tang et al. 2022) aims to discover the spatial correspondence between content images and style images to learn the local style details, but its reference selection strategy is sensitive to the quality of results. CF-Font (Wang et al. 2023) fuses various content features of different fonts and introduces an iterative style-vector refinement strategy. However, these methods still struggle with generating complex characters and handling large style variations.

Diffusion Model

Recently, diffusion models have achieved rapid development in vision generation tasks. Several prominent conditional diffusion models have been developed (Nichol et al. 2021; Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Zhang and Agrawala 2023; Ruiz et al. 2023). LDM (Rombach et al. 2022) proposes a cross-attention mechanism to incorporate the condition into the UNet and treats the diffusion process in the latent space. In text image generation, (Luhman and Luhman 2020; Gui et al. 2023; Nikolaidou et al. 2023) apply diffusion models to generate handwritten characters and demonstrate their promising effects. CTIG-DM (Zhu et al. 2023) devises image, text, and style as conditions and introduces four text image generation modes in a diffusion model. In contrast to general image generation, font generation requires distinct stroke details and intricate structural features at a fine-grained level. This motivates us to harness multi-scale content features and propose an innovative style contrastive learning strategy.

Methodology

As shown in Figure 2, our proposed method consists of a Conditional Diffusion model and a Style Contrastive Refinement module. In the **Conditional Diffusion** model, given
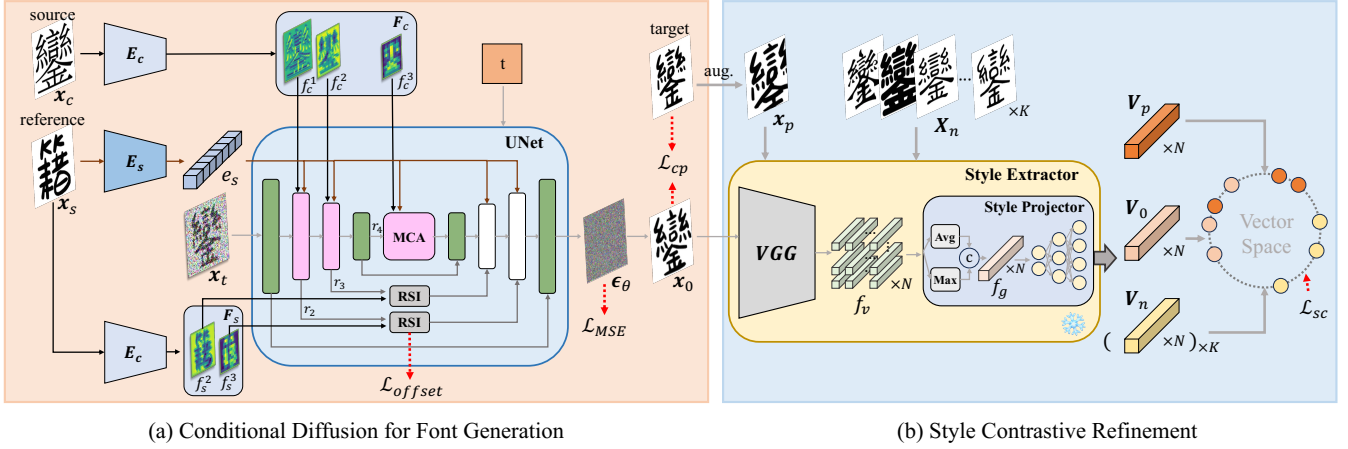
(a) Conditional Diffusion for Font Generation   (b) Style Contrastive Refinement

Figure 2: Overview of our proposed method. (a) The Conditional Diffusion model is a UNet-based network composed of a content encoder E_c and a style encoder E_s. The reference image x_s is passed through a style encoder E_s and a content encoder E_c respectively, obtaining a style embedding e_s and structure maps F_s. The source image is encoded by a content encoder E_c. To obtain multi-scale features F_c, we derive output from the different layers of E_c and inject each of them through our proposed MCA block. RSI block is employed to conduct spatial deformation from reference structural features F_s. (b) The Style Contrastive Refinement module is to disentangle different styles from images and provide guidance to the diffusion model.

a source image x_c and a reference image x_s, our goal is to train a conditional diffusion model where the final output image should not only have the same content as in x_c, but should also be consistent with the reference style. **Style contrastive refinement** module aims to disentangle different styles from a group of images and offer guidance to the diffusion model via a style contrastive loss.

Conditional Diffusion for Font Generation

Based on DDPM (Ho, Jain, and Abbeel 2020), the general idea of our diffusion-based image-to-image font generation method is to design a forward process that incrementally adds noise to the target distributions $x_0 \sim q(x_0)$, while the denoising process involves learning the reverse mapping. The denoising process aims to transform a noise $x_T \sim (0, I)$ to the target distribution in T steps.

Specifically, the forward process of FontDiffusers is a Markov chain and the noise adding process can be summarized as follows:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $t \sim [0, T]$, ϵ is the added Gaussian noise. $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=0}^{t}(1 - \beta_i)$, $\beta_i \sim (0, 1)$ is a fixed hyper-parameter of variance. During the reverse process, the reverse mapping can be approximated by a model to predict the noise $\epsilon_\theta(x_t, t, x_c, x_s)$ and then obtain the x_{t-1} as follows:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(x_t, t, x_c, x_s)) + \sigma_t z, \tag{2}$$

where σ_t is the hyper-parameter and noise $z \sim (0, I)$.

We predict the noise $\epsilon_\theta(x_t, t, x_c, x_s)$ using our conditional diffusion model. Specifically, to enhance the preservation of complex characters, we employ a *Multi-scale Content Aggregation* (MCA) block to inject the global and local content cues into the UNet of our model. Moreover, a

Reference-Structure Interaction (RSI) block is employed to facilitate structural deformation from the reference features.

Multi-Scale Content Aggregation (MCA) Generating complex characters has always been challenging, and many existing methods only rely on a single-scale content feature, disregarding the intricate details such as strokes and components. As shown in Figure 3, large-scale features retain lots of detailed information while small-scale are lacking.
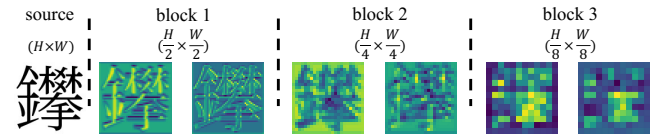

Figure 3: Content features in various blocks.

Therefore, we employ a Multi-scale Content Aggregation (MCA) block, injecting global and local content features across different scales into the UNet of our diffusion model. Specifically, the source image x_c is first embedded by the content encoder E_c, obtaining multi-scale content features $F_c = \{f_c^1, f_c^2, f_c^3\}$ from different layers. Together with the style embedding e_s encoded by the style encoder E_s, each content feature f_c^i is injected into the UNet through three MCA modules respectively. As illustrated in Figure 4, the content feature f_c^i is concatenated with the previous UNet block feature r_i, resulting in a channel-informative feature I_c. To enhance the capability of adaptive selective channel fusion, we apply a channel attention (Hu, Shen, and Sun 2018) on I_c, in which an average pooling, two 1×1 convolutions and an activation function are employed. The attention results in a global channel-aware vector W_c, which is used to weight the channel-informative feature I_c via channel-wise
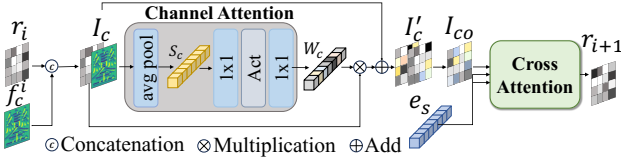
Figure 4: Multi-scale Content Aggregation.

multiplication. Then, after a residual connection, we employ a 1×1 convolution to reduce the channel number of I_c', obtaining the output I_{co}. Lastly, we apply a cross-attention module to insert the style embedding e_s, in which e_s is employed as Key and Value, while I_{co} is employed as Query.

Reference-Structure Interaction (RSI) There exists structural differences (*e.g.*, font size) between the source image and the target image. To address this issue, we propose a Reference-Structure Interaction (RSI) block that employs deformable convolutional networks (DCN) (Dai et al. 2017) to conduct structural deformation on the skip connection of UNet. In contrast to (Xie et al. 2021), our conditional model directly extracts structural information from the reference features to obtain the deformation offset δ_{offset} for DCN.

Specifically, the reference image $\boldsymbol{x}_s$ is first passed through the content encoder E_c to obtain the structure maps $\boldsymbol{F}_s = \{f_s^1, f_s^2\}$, and each f_s^i is as the input to both RSI modules respectively. Due to the misalignment in the spatial position between the UNet feature and the reference feature, we introduce a cross-attention to enable long-distance interactions to obtain the offset δ_{offset}, instead of CNN. The interaction process can be summarized in Equation 3: r_i is the UNet feature. And the essential element of this process involves leveraging the UNet feature r_i and structure map f_s^i in a softmax operation, which primarily calculates the region of interest relative to each query position.

$$S_s \in \mathbb{R}^{C_f^i \times H_i W_i} = flatten(f_s^i),$$
$$S_r \in \mathbb{R}^{C_r^i \times H_i W_i} = flatten(r_i),$$
$$Q = \Phi_q(S_s), \quad K = \Phi_k(S_r), \quad V = \Phi_v(S_r),$$
$$F_{attn} = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad \delta_{offset} = FFN(F_{attn}),$$
$$I_R = DCN(r_i, \delta_{offset}), \tag{3}$$

where Φ_q, Φ_k, Φ_v are linear projections, and FFN denotes the feed forward network. I_R is the output of RSI.

Style Contrastive Refinement

One purpose of font generation is to achieve the intended style imitating effect, regardless of the variations of style between the source and the reference. A novel strategy is to find a suitable style representation and further provide feedback to our model. Therefore, we propose a Style Contrastive Refinement (SCR) module, a font style representation learning module that disentangles style from a group of samples images and incorporates a style contrastive loss to supervise our diffusion model, ensuring the generated style aligns with the target at the global and local level.

The architecture of SCR is shown on the right of Figure 2, which consists of a style extractor. Inspired by (Zhang et al. 2022), a VGG network is employed to embed the font image in the extractor. To capture both global and local style characteristics effectively, we select N layers of feature maps $\boldsymbol{F}_v = \{f_v^0, f_v^1, ..., f_v^N\}$ from VGG network, utilizing them as input to a style projector. The projector applies an average pooling and a maximum pooling to extract different global channel features separately, and then concatenates both of them channel-wise, resulting in the features $\boldsymbol{F}_g = \{f_g^0, f_g^1, ..., f_g^N\}$. Finally, after several linear projections, style vectors $\boldsymbol{V} = \{v^0, v^1, ..., v^N\}$ are obtained.

The style vectors $\boldsymbol{V}$ can provide supervising signals to the diffusion model and guide it to imitate style. Therefore, we adopt a contrastive learning strategy, in which we leverage a pre-trained SCR and incorporate a style contrastive loss $\mathcal{L}_{sc}$ to supervise whether the style of the generated sample $\boldsymbol{x}_0$ is consistent with the target style and distinguishable from negative styles. To ensure content-irrelevance and style-relevance, we choose the target image as the positive sample and select K negative samples that are with different styles but the same content, rather than directly considering the rest of the chosen target sample as negatives. Therefore, the supervision of SCR can be summarized as follows:

$$\boldsymbol{V}_0 = Extrac(\boldsymbol{x}_0), \quad \boldsymbol{V}_p = Extrac(\boldsymbol{x}_p), \quad \boldsymbol{V}_n = Extrac(\boldsymbol{x}_n)$$
$$\mathcal{L}_{sc} = -\sum_{l=0}^{N-1} log \frac{exp(v_0^l \cdot v_p^l / \tau)}{exp(v_0^l \cdot v_p^l / \tau) + \sum_{i=1}^{K} exp(v_0^l \cdot v_{n_i}^l / \tau)}, \tag{4}$$

where $Extrac$ represents the style extractor. K is the number of negative samples. $\boldsymbol{V}_0$, $\boldsymbol{V}_p$ and $\boldsymbol{V}_n$ denote the style vectors of generated, positive and negative samples respectively, and $v_0^l, v_p^l, v_{n_i}^l$ denotes the l-th generated, positive and negative layer vector respectively. τ is a temperature hyperparameter and set as 0.07. The pre-training details of SCR are listed in Appendix.

To enhance the robustness of style imitation, we apply an *augmentation strategy* on the positive target sample, which includes random cropping and random resizing.

Training Objective

Our training adopts a coarse-to-fine two-phase strategy.

Phase 1 During phase 1, we optimize FontDiffuser mainly with the standard MSE diffusion loss, excluding the SCR module. This ensures that our generator acquires the fundamental capability for font reconstruction:

$$\mathcal{L}_{total}^1 = \mathcal{L}_{MSE} + \lambda_{cp}^1 \mathcal{L}_{cp} + \lambda_{off}^1 \mathcal{L}_{offset}, \tag{5}$$

in which,

$$\mathcal{L}_{MSE} = \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{x}_c, \boldsymbol{x}_s) \|^2, \tag{6}$$
$$\mathcal{L}_{cp} = \sum_{l=1}^{L} \| \mathcal{VGG}_l(\boldsymbol{x}_0) - \mathcal{VGG}_l(\boldsymbol{x}_{target}) \|, \tag{7}$$
$$\mathcal{L}_{offset} = mean(\| \delta_{offset} \|), \tag{8}$$

where $\mathcal{L}_{total}^1$ represents the total loss in phase 1. $\mathcal{VGG}_l(\cdot)$ is the layer feature encoded by VGG and L is the number of the

Model		Easy				Medium				Hard				Average				User (%)
		FID	SSIM	LPIPS	L1	FID	SSIM	LPIPS	L1	FID	SSIM	LPIPS	L1	FID	SSIM	LPIPS	L1	
SFUC	FUNIT	11.34	.4342	.1985	.3888	11.02	.3516	.2144	.4474	17.91	.3271	.2374	.4648	9.67	.3755	.2146	.4305	1.30
	LF-Font	18.01	.4914	.1770	.3325	25.72	.3833	.2048	.4184	39.68	.3444	.2343	.4511	22.74	.4127	.2024	.3955	0.37
	DG-Font	20.48	.4613	.2111	.3610	24.44	.3831	.2354	.4146	29.60	.3444	.2614	.4430	21.16	.4016	.2333	.4024	5.28
	MX-Font	12.43	.4693	.1688	.3511	11.29	.3790	.1784	.4184	14.11	.3338	.1964	.4546	10.22	.4002	.1796	.4033	11.39
	Fs-Font	27.20	.4282	.2258	.3869	27.94	.3425	.2394	.4536	36.90	.3091	.2621	.4800	25.99	.3651	.2404	.4361	0.83
	CG-GAN	**8.23**	.4692	.1816	.3582	**9.11**	.3755	.1952	.4280	14.19	.3396	.2173	.4570	7.79	.4004	.1961	.4100	32.87
	CF-Font	14.08	.4924	.2015	.3224	13.96	.4006	.2347	.3929	16.84	.3662	.2634	.4197	12.13	.4253	.2301	.3741	1.11
	Ours	8.51	**.5370**	**.1316**	**.2901**	9.46	**.4462**	**.1411**	**.3623**	**11.15**	**.4033**	**.1562**	**.3986**	**7.70**	**.4682**	**.1416**	**.3454**	46.85
UFUC	FUNIT	14.55	.4507	.1839	.3720	16.09	.3495	.2045	.4484	25.97	.2963	.2403	.4918	13.14	.3655	.2095	.4374	2.03
	LF-Font	23.92	.4949	.1687	.3301	38.61	.3746	.1997	.4257	55.44	.3071	.2370	.4833	32.89	.3922	.2018	.4130	.10
	DG-Font	25.61	.4788	.1957	.3450	27.08	.3803	.2172	.4165	32.73	.3254	.2421	.4561	22.71	.3948	.2183	.4059	8.99
	MX-Font	14.92	.4808	.1552	.3408	14.09	.3786	.1625	.4195	16.40	.3189	.1783	.4689	10.77	.3928	.1653	.4098	14.11
	Fs-Font	42.78	.4524	.2100	.3646	43.69	.3495	.2282	.4448	49.33	.2973	.2565	.4869	38.77	.3664	.2315	.4321	1.26
	CG-GAN	14.14	.4887	.1677	.3369	14.41	.3793	.1831	.4173	26.89	.3114	.2120	.4710	12.93	.3931	.1876	.4084	20.97
	CF-Font	22.09	.4841	.1901	.3322	24.58	.3897	.2180	.4046	25.83	**.3434**	.2461	**.4420**	19.69	.4057	.2180	.3929	5.41
	Ours	**12.90**	**.5080**	**.1418**	.3175	**11.63**	**.4117**	**.1468**	.3926	13.12	.3420	**.1600**	.4508	**8.54**	**.4206**	**.1496**	**.3870**	47.15

Table 1: Quantitative Results on SFUC and UFUC. 'User' denotes the user study. 'Average' and the user study is evaluated on all characters of three levels of complexity. The bold indicates the state-of-the-art and the underline indicates the second best.

chosen layers. $\mathcal{L}_{cp}$ is used to penalize the content misalignment between generated VGG features of x_0 and the corresponding x_{target} target features. The offset loss $\mathcal{L}_{offset}$ is used to constrain the offset in our RSI module and $mean$ is the averaging process. $\lambda^1_{cp} = 0.01$ and $\lambda^1_{off} = 0.5$.

Phase 2 In phase 2, we implement the SCR module, incorporating the style contrastive loss, to provide style imitation guidance to the diffusion model at the global and local levels. Thus our conditional diffusion model in phase 2 is optimized by:

$$\mathcal{L}^2_{total} = \mathcal{L}_{MSE} + \lambda^2_{cp}\mathcal{L}_{cp} + \lambda^2_{off}\mathcal{L}_{offset} + \lambda^2_{sc}\mathcal{L}_{sc}, \quad (9)$$

where $\mathcal{L}^2_{total}$ represents the total loss in phase 2. The hyperparameters $\lambda^2_{cp} = 0.01$, $\lambda^2_{off} = 0.5$ and $\lambda^2_{sc} = 0.01$.

Experiment

Datasets and Evaluation Metrics

We collect a Chinese font dataset of 424 fonts. We randomly select 400 fonts (referred to as "seen fonts") with 800 Chinese characters (referred to as "seen characters") as training set. We evaluate methods on two test sets: one includes 100 randomly selected seen fonts, which contains 272 characters that were not seen during training (referred to as "SFUC"), and the other test set consists of 24 unseen fonts and 300 unseen characters (referred to as "UFUC"). The categorization details of three levels of complexity are in Appendix. Moreover, we additionally conduct a comparison on 24 unseen fonts and 800 seen characters (referred to as "UFSC").

For quantitative evaluation, we adopt FID, SSIM, LPIPS, and L1 loss metrics. Pixel-level metrics SSIM and L1 loss are employed to measure the per-pixel consistency. LPIPS (Zhang et al. 2018) and FID (Heusel et al. 2017) are perceptual metrics, which are closer to human visual perception.

Furthermore, we conduct a *user study* to assess the subjective quality of images. In total, 25 participants are asked for the evaluation and the results are presented in Table 1.

Implementation Details

We train FontDiffuser using AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The image size is set as 96. Moreover, following (Ho and Salimans 2022), we simply drop out the source image and the reference image with the probability of 0.1. In phase 1, we train the model with a batch size of 16 and a total step of 440000. The learning rate is set as $1e-4$ with linear schedule. In phase 2, the learning rate is set as $1e-5$ and is fixed as constant. We train with a batch size of 16, a total step of 30000, and negative samples of 16. The experiments are conducted on a single RTX 3090 GPU.

During sampling, we adopt a classifier-free guidance strategy (Ho and Salimans 2022) to amplify the effect of the conditions x_c and x_s. We set the unconditional content image and unconditional style image to pixel 255 as $\emptyset$, and our sampling strategy can be formulated as:

$$\epsilon_\theta(x_t, t, x_c, x_s) = (1-s)\epsilon_\theta(x_t, t, \emptyset, \emptyset) + s\epsilon_\theta(x_t, t, x_c, x_s), \quad (10)$$

where s is the guidance scale and is set as 7.5 in the experiments. To speed up sampling, we use the DPM-Solver++ sampler (Lu et al. 2022) with only 20 inference steps.

Comparison with State-of-the-Art Method

We compare our method with seven state-of-the-art methods: one image-to-image translation method (FUNIT (Liu et al. 2019)) and six Chinese font generation methods (LF-Font (Park et al. 2021a), MX-Font (Park et al. 2021b), DG-Font (Xie et al. 2021), CG-GAN (Kong et al. 2022), Fs-Font (Tang et al. 2022), and CF-Font (Wang et al. 2023)). Additionally, we compare with Diff-Font (He et al. 2022) on Unseen Font Seen Character (UFSC). For a fair comparison,

Model	FID	SSIM	LPIPS	L1
LF-Font	18.67	.4823	.1688	.3400
DG-Font	19.81	.4532	.2047	.3646
MX-Font	9.32	.4605	.1603	.3571
Fs-Font	31.40	.4270	.2160	.3855
CG-GAN	7.72	.4655	.1721	.3544
CF-Font	14.20	.4396	.2139	.3713
Diff-Font	12.08	.4192	.2022	.3877
Ours	**7.67**	**.4942**	**.1426**	**.3279**

Table 2: Quantitative Results on UFSC.

Figure 5: Cross-lingual generation (Chinese to Korean).

Module			FID	SSIM	LPIPS	L1
M	R	S				
✗	✗	✗	8.12	.4112	.1526	.3955
✓	✗	✗	**7.84**	.4114	.1511	.3954
✓	✓	✗	8.44	.4137	.1506	.3925
✓	✓	✓	8.54	**.4206**	**.1496**	**.3870**

Table 3: Effectiveness of different modules. M, R, and S represent MCA, RSI, and SCR respectively. The first row represents the baseline.

Figure 6: Visualization of different modules. M, R, and S represent MCA, RSI, and SCR respectively. Red boxes represent the missing strokes while green represents the corresponding improvements. Blue denotes structural promotion.

we use the font of Song as the source, and all methods are trained based on their official codes.

Quantitative Comparison The quantitative results are presented in Table 1. FontDiffuser achieves the best performance across all matrices at average level, showing a significant gap compared to other methods on both SFUC and UFUC. It indicates that FontDiffuser can generate fonts that are visually closer to human perception. At easy and medium levels, though FID in SFUC ranks second, FontDiffuser outperforms other methods in the remaining metrics, particularly the perceptual matrix LPIPS. At hard level, our method performs the best in SFUC and achieves the best FID and LPIPS scores in UFUC. It should be noted that SSIM and L1 loss are pixel-level metrics, which may not directly reflect the overall performance. For instance, an impressive visual result may not perfectly match the target pixel to pixel. The hard-level results demonstrate the advantage of FontDiffuser in generating complex characters. Furthermore, as shown in Table 2, FontDiffuser achieves state-of-the-art performance on UFSC. Notably, Diff-Font (He et al. 2022) is only capable of generating seen characters, and our method also outperforms it by a significant margin.

Qualitative Comparison In Figure 7, we provide visualizations of the results on SFUC and UFUC, which intuitively reflect the visual effects of different methods. FontDiffuser consistently generates high-quality results and performs better in terms of content preservation, style consistency, and structural correctness compared with other state-of-the-art methods. Particularly, our method demonstrates significant superiority in generating complex characters and handling large variations in style transfer, while other methods still exhibit issues such as missing strokes, artifacts,

blurriness, layout errors, and style inconsistency. We also present some cross-lingual generation samples (Chinese to Korean) in Figure 5, which are generated by our method. It demonstrates that FontDiffuser is flexible in generating for other languages and exhibits cross-domain capability though our model is trained by Chinese dataset.

Ablation Studies

In this section, we conduct several ablation studies to analyze the performance of our proposed modules and strategies. The experiments are tested on the unseen font unseen characters (UFUC) at average level.

Effectiveness of Different Modules We separate the proposed MCA, RSI, and SCR, and progressively add them to the baseline. The baseline concatenates the content image with x_t as the input of UNet. Table 3 shows that the quantitative results of these three modules are improved in terms of SSIM, LPIPS, and L1 loss, except for FID. Additionally, these modules also contribute to visual enhancements, as shown in Figure 6. For example, in the first row of Figure 6, the issue of missing strokes in the baseline is mitigated by the incorporation of the MCA module.

Effectiveness of Augmentation Strategy in SCR We investigate the advantage of the proposed augmentation strategy in SCR, in which FontDiffuser is trained with and without augmentation strategy during the training phase 2. As shown in Table 4, it clearly demonstrates that the augmentation strategy boosts the generation performance in terms of SSIM, LPIPS, and L1 loss.

Comparison between Cross-attention Interaction and CNN in RSI We conduct a comparative analysis between

Figure 7: Qualitative comparison on SFUC and UFUC. Red boxes highlight the failures of other methods.

Method	FID	SSIM	LPIPS	L1
w/o augmentation	**8.18**	.4172	.1504	.3900
augmentation	8.54	**.4206**	**.1496**	**.3870**

Table 4: Effectiveness of augmentation strategy in SCR.

Method	FID	SSIM	LPIPS	L1
CNN	9.17	.4130	.1537	.3932
cross-attention	**8.54**	**.4206**	**.1496**	**.3870**

Table 5: Comparison between cross-attention and CNN.

cross-attention interaction and CNN interaction in RSI. The results in Table 5 show that the cross-attention interaction in RSI outperforms the CNN-based in all matrices, showcasing the superiority of our proposed method.

Visualization of SCR Contrastive Score

We provide visualization of the SCR contrastive score in Figure 8, which demonstrates that SCR can effectively distinguish the target from a group of samples, even though some of them exhibit similar styles. By combining SCR with style contrastive loss, we observe that SCR can refine the generated style through a learning-by-contrast manner.

Figure 8: Visualization of SCR contrastive score. The left column represents the generated samples. Each row corresponds to the chosen samples. Red boxes highlight the target while blues highlight samples similar to the generated style.

Conclusion

In this paper, we propose a diffusion-based image-to-image font generation method, called FontDiffuser, which excels in generating complex characters and handling large variations in style transfer. Specifically, we propose the MCA block to inject multi-scale content features into our diffusion model, enhancing the preservation of complex characters. Moreover, we propose a novel style representation learning strategy, which implements the SCR module and uses a style contrastive loss to supervise our diffusion model. Additionally, an RSI block is employed to facilitate structural deformation using reference features. Extensive experiments demonstrate that FontDiffuser outperforms the state-of-the-art method on characters of three levels of complexity. Furthermore, FontDiffuser demonstrates its applicability to the cross-lingual font generation task (*e.g.*, Chinese to Korean), highlighting its promising cross-domain capability.

Acknowledgements

References

Cha, J.; Chun, S.; Lee, G.; Lee, B.; Kim, S.; and Lee, H. 2020. Few-shot compositional font generation with dual memory. In *Proc. ECCV*, 735–751. Springer.

Chang, J.; Gu, Y.; Zhang, Y.; Wang, Y.-F.; and Innovation, C. 2018. Chinese Handwriting Imitation with Hierarchical Generative Adversarial Network. In *Proc. BMVC*, 290.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proc. ICCV*, 764–773.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Proc. NeurIPS*, 27.

Gui, D.; Chen, K.; Ding, H.; and Huo, Q. 2023. Zero-shot Generation of Training Data with Denoising Diffusion Probabilistic Model for Handwritten Chinese Character Recognition. *arXiv preprint arXiv:2305.15660*.

He, H.; Chen, X.; Wang, C.; Liu, J.; Du, B.; Tao, D.; and Qiao, Y. 2022. Diff-Font: Diffusion Model for Robust One-Shot Font Generation. *arXiv preprint arXiv:2212.05895*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Proc. NeurIPS*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Proc. NeurIPS*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proc. CVPR*, 7132–7141.

Huang, Y.; He, M.; Jin, L.; and Wang, Y. 2020. RD-GAN: Few/zero-shot chinese character style transfer via radical decomposition and rendering. In *Proc. ECCV*, 156–172. Springer.

Jiang, Y.; Lian, Z.; Tang, Y.; and Xiao, J. 2017. DCFont: an end-to-end deep Chinese font generation system. In *SIGGRAPH Asia Technical Briefs*, 1–4.

Kong, Y.; Luo, C.; Ma, W.; Zhu, Q.; Zhu, S.; Yuan, N.; and Jin, L. 2022. Look closer to supervise better: one-shot font generation via component-based discriminator. In *Proc. CVPR*, 13482–13491.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. In *Proc. ICCV*, 10551–10560.

Liu, W.; Liu, F.; Ding, F.; He, Q.; and Yi, Z. 2022. XMP-Font: self-supervised cross-modality pre-training for few-shot font generation. In *Proc. CVPR*, 7905–7914.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*.

Luhman, T.; and Luhman, E. 2020. Diffusion models for handwriting generation. *arXiv preprint arXiv:2011.06704*.

Lyu, P.; Bai, X.; Yao, C.; Zhu, Z.; Huang, T.; and Liu, W. 2017. Auto-encoder guided GAN for Chinese calligraphy synthesis. In *Proc. ICDAR*, volume 1, 1095–1100. IEEE.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nikolaidou, K.; Retsinas, G.; Christlein, V.; Seuret, M.; Sfikas, G.; Smith, E. B.; Mokayed, H.; and Liwicki, M. 2023. WordStylist: Styled Verbatim Handwritten Text Generation with Latent Diffusion Models. *arXiv preprint arXiv:2303.16576*.

Park, S.; Chun, S.; Cha, J.; Lee, B.; and Shim, H. 2021a. Few-shot font generation with localized style representations and factorization. In *Proc. AAAI*, volume 35, 2393–2402.

Park, S.; Chun, S.; Cha, J.; Lee, B.; and Shim, H. 2021b. Multiple heads are better than one: Few-shot font generation with multiple localized experts. In *Proc. ICCV*, 13900–13909.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Proc. NeurIPS*, 35: 36479–36494.

Sun, D.; Ren, T.; Li, C.; Su, H.; and Zhu, J. 2017. Learning to write stylized chinese characters by reading a handful of examples. *arXiv preprint arXiv:1712.06424*.

Sun, D.; Zhang, Q.; and Yang, J. 2018. Pyramid embedded generative adversarial network for automated font generation. In *Proc. ICPR*, 976–981. IEEE.

Tang, L.; Cai, Y.; Liu, J.; Hong, Z.; Gong, M.; Fan, M.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Few-shot font generation by learning fine-grained local styles. In *Proc. CVPR*, 7895–7904.

Tian, Y. 2017. zi2zi: Master Chinese Calligraphy with Conditional Adversarial Networks. http://github.com/kaonashi-tyc/zi2zi.

Wang, C.; Zhou, M.; Ge, T.; Jiang, Y.; Bao, H.; and Xu, W. 2023. CF-Font: Content Fusion for Few-shot Font Generation. In *Proc. CVPR*, 1858–1867.

Wu, S.-J.; Yang, C.-Y.; and Hsu, J. Y.-j. 2020. CalliGAN: Style and structure-aware Chinese calligraphy character generator. *arXiv preprint arXiv:2005.12500*.

Xie, Y.; Chen, X.; Sun, L.; and Lu, Y. 2021. DG-Font: Deformable generative networks for unsupervised font generation. In *Proc. CVPR*, 5130–5140.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 586–595.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2022. Domain enhanced arbitrary image style transfer via contrastive learning. In *Proc. SIGGRAPH*, 1–8.

Zhang, Y.; Zhang, Y.; and Cai, W. 2018. Separating style and content for generalized style transfer. In *Proc. CVPR*, 8447–8455.

Zhu, Y.; Li, Z.; Wang, T.; He, M.; and Yao, C. 2023. Conditional Text Image Generation with Diffusion Models. In *Proc. CVPR*, 14235–14245.