One-DM: One-Shot Diffusion Mimicker for Handwritten Text Generation

Gang Dai[1⋆], Yifan Zhang[2,3⋆],
Quhui Ke[1], Qiangya Guo[1], and Shuangping Huang[1,4†]

[1]South China University of Technology
[2]National University of Singapore, [3]Skywork AI, [4]Pazhou Laboratory
{eedaigang@mail., eehsp@}scut.edu.cn, yifan.zhang@u.nus.edu

Abstract. Existing handwritten text generation methods often require more than ten handwriting samples as style references. However, in practical applications, users tend to prefer a handwriting generation model that operates with just a single reference sample for its convenience and efficiency. This approach, known as "one-shot generation", significantly simplifies the process but poses a significant challenge due to the difficulty of accurately capturing a writer's style from a single sample, especially when extracting fine details from the characters' edges amidst sparse foreground and undesired background noise. To address this problem, we propose a One-shot Diffusion Mimicker (One-DM) to generate handwritten text that can mimic any calligraphic style with only one reference sample. Inspired by the fact that high-frequency information of the individual sample often contains distinct style patterns (*e.g.*, character slant and letter joining), we develop a novel style-enhanced module to improve the style extraction by incorporating high-frequency components from a single sample. We then fuse the style features with the text content as a merged condition for guiding the diffusion model to produce high-quality handwritten text images. Extensive experiments demonstrate that our method can successfully generate handwriting scripts with just one sample reference in multiple languages, even outperforming previous methods using over ten samples. Our source code is available at https://github.com/dailenson/One-DM.

Keywords: Handwritten Text Generation · One-Shot Generation

1 Introduction

In the digital age, handwriting text generation blends the personalization of traditional handwriting with the efficiency of automated processes, offering a digital format to preserve the authenticity of individual handwriting. This task aims to automatically generate the desired handwritten text images that not only correspond to specific text content, but also emulate the calligraphic style of a given exemplar writer (e.g., character slant, cursive join, stroke thickness, and

⋆ Authors contributed equally; † Corresponding author

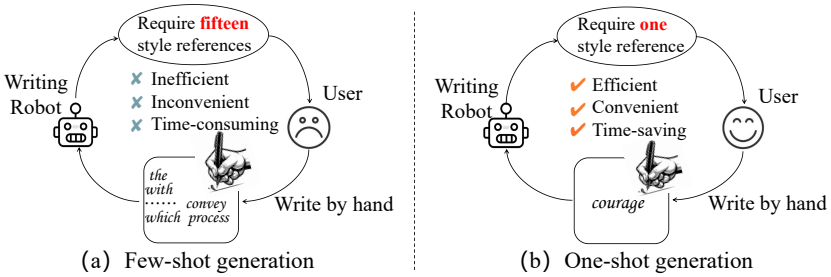


Fig. 1: User experience comparisons between one-shot and few-shot handwritten text generation methods. It reveals that one-shot setting leads to a better user experience.

ink color). This provides great convenience for people with hand impairments and also contributes to accelerating the process of handwritten font design.

Previous works [4, 8, 11–13, 24, 39] introduce generative adversarial networks (GANs) [5,14,32] for handwritten text generation. For instance, ScrabbleGAN [11] leverages random noises as style inputs and conditions content inputs on character-level labels, enabling the synthesis of handwritten words with randomly sampled styles. In a recent study [9], diffusion models [31, 42, 53] such as denoising diffusion probabilistic models (DDPM) [18] have demonstrated even higher quality of image generation compared to GANs. This motivates several attempts, such as WordStylist [37], GC-DDPM [10], and CTIG-DM [57], to condition the denoising process on the fixed writer ID to generate handwritten text images with controllable styles. However, the major limitation of these methods [10,11,37,57] is that they are unable to mimic the unseen writers' handwriting styles.

To imitate any given handwriting style, some previous methods [4, 24, 39], known as few-shot generation, require users to provide a few samples (typically 15) as style references. They employ a style encoder to extract writing styles from the given samples, thus offering flexible control over the style of the generated samples. However, the traditional few-shot generation pipeline is inconvenient, inefficient, and time-consuming, as shown in Figure 1. Users prefer methods that only require a single sample as a style reference, known as one-shot generation methods, because they are more convenient, efficient, and easy-to-use. Our goal is to investigate the more challenging one-shot generation task that holds significant practical value, striving to achieve high-quality handwritten text image generation with desired styles and contents.

The primary challenge of one-shot generation task lies in accurately extracting a user's handwriting style from just one style reference image. As illustrated in Figure 2, characters, being abstract symbols, typically occupy only a minor portion of the reference image. Besides, the reference image is often cluttered with noisy background information, which poses a significant obstacle in extracting the individual handwriting style. Previous one-shot generation methods [8,12,13] simply follow the architectural design of few-shot approaches, using a vanilla CNN encoder to directly extract the handwriting style from a single sample. The extracted style is then combined with text contents and input into a CNN decoder to generate the desired handwritten images. These methods ex-

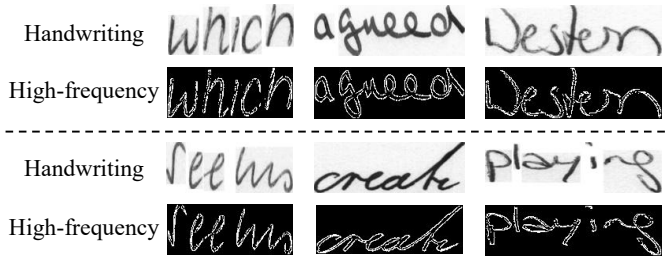


Fig. 2: Handwritten text samples and corresponding high-frequency components. We find that high-frequency components have more pronounced character contours, clearly showcasing the style patterns, such as character slant and cursive connections.

hibit a limited performance in emulating handwriting styles, due to their poor ability of style extraction.

To address the above challenges, our key idea revolves around utilizing the high-frequency information of the sample to enhance the extraction of the handwriting style. As depicted in Figure 2, the high-frequency information encompasses the overall contours of the handwritten text, allowing for a clear observation of key style patterns such as the text slant, letter spacing, and cursive connections. Hence, the incorporation of high-frequency information facilitates a more effective handwriting style extraction.

In light of the above insight, we propose a One-shot Diffusion Mimicker (One-DM) for handwritten text generation, which is simultaneously guided by desired style and arbitrary content. More specifically, we first develop a style-enhanced module to process both a style reference image and its high-frequency components in parallel. Considering that the reference image often contains background noise, we design a gate mechanism to suppress the inflow of background noise. Regarding the high-frequency components, such as character slant and ligature that show clearer style patterns, we employ a contrastive learning framework [25,54,55] to further obtain discriminative style features, guiding the handwritten text synthesis with both realistic and diverse styles. Subsequently, the style features extracted from both branches are adaptively fused with the specific content prototype in a style-content fusion module. Finally, the seamlessly integrated style and content features act as conditional inputs, guiding the denoising process for the progressive synthesis of stylized handwritten text images. It is worth noting that our One-DM effectively mimics a user's writing style with only one reference sample through our style-enhanced module, surpassing few-shot methods in producing higher-quality stylized handwritings.

To sum up, our contributions are as follows:

- We propose a novel diffusion model for stylized handwritten text generation, which only requires a single reference sample as style input, and imitates its writing style to generate handwritten text with arbitrary content.
- We introduce the high-frequency components of the reference sample to enhance the extraction of handwriting style. The proposed style-enhanced

module can effectively capture the writing style patterns and suppress the interference of background noise.
– Extensive experiments on handwriting datasets in English, Chinese, and Japanese demonstrate that our approach with a single style reference even outperforms previous methods with 15x-more references.

2 Related Work

Handwriting generation. Handwritten text is typically stored in two formats: online trajectory form or offline image form. Online handwriting generation methods [1, 7, 28, 44, 46, 56] often employ Recurrent Neural Networks (RNNs) [1, 6, 28, 44–46, 56], transformer decoders [7], or diffusion models [34, 40] to progressively generate writing trajectories. Unlike online methods, offline generation methods have the advantage of not requiring additional trajectory supervision information and can generate realistic handwritten text with stroke width and ink color, which online methods cannot produce.

Previous offline handwriting generation methods [2, 4, 8, 11, 20, 24, 35, 39] in deep learning predominantly rely on Generative Adversarial Networks (GANs). Early works [2, 11] condition the generative process on the word embeddings [2] or concatenated letter-tokens [11] to synthesize handwritten word images. However, these methods struggle to flexibly control the writing style. Thus, few-shot approaches [4, 24, 39] which rely on 15 style references are introduced. For instance, GANwriting [24] utilizes a CNN encoder [16, 50, 51] to extract a user's handwriting style from a few samples, which are then combined with specific text content to generate handwritings in the desired style. In its follow-up work [23], the synthesized samples are demonstrated to assist in training more robust handwritten text recognizers. Further, HWT [4] employs a transformer encoder [29, 47] to extract rich style patterns from reference samples, enhancing the performance in style mimicry. Recently, VATr [39] uses images of symbols as content representations, enabling the generation of out-of-charset characters.

Concurrently, one-shot generation methods [8, 12, 13] are proposed. Despite their ability to mimic handwriting styles with only a single sample, these techniques still lag behind few-shot methods in terms of stylized generation results. Additionally, previous handwriting generation methods [2, 4, 8, 11–13, 24] often rely on complex content representations, such as recurrent embeddings [2] and letter-level tokens [4, 8, 11–13, 24]. SLOGAN [35] extracts textual contents from easily obtainable printed images. However, it faces challenges in generalizing to unseen writing styles due to its fixed writer ID. Similarly, some diffusion-based [10, 37, 57] methods condition the denoising process on fixed style ID and are unable to mimic styles that they have not previously encountered. In contrast, our One-DM effectively obtains style information from one style sample and thus can generate handwritings with arbitrary styles. Due to space constraints, we discuss more related works in supplementary, including diffusion methods for general image generation, contrastive-based method [48], and frequency-based methods [30, 38].

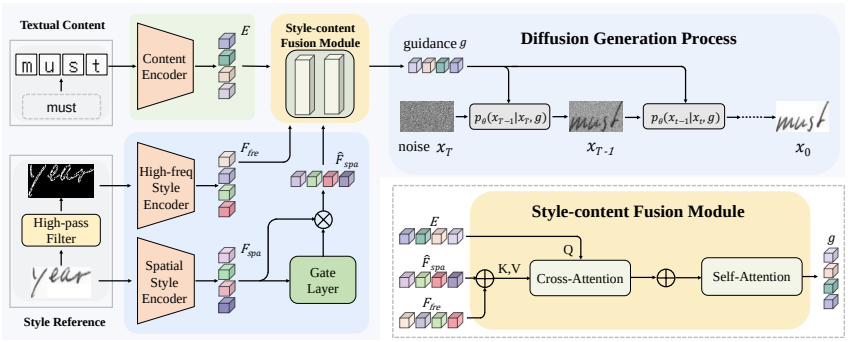


Fig. 3: Overview of the proposed method. The style reference initially passes through a high-pass filter to extract its high-frequency components. Subsequently, the spatial and the high-frequency style encoders independently extract style features F_{spa} and F_{fre} from the style reference and its high-frequency information, respectively. F_{spa}, after being filtered through a gate mechanism, is fused with F_{fre} and content features E in the fusion module. The merged feature then serves as a condition input to guide the diffusion generation process.

3 Method

Problem statement. We aim to synthesize handwritten text images that are controlled by both textual content and handwriting style. Given any word string $\mathcal{A}$ and a single style sample I_s from a writer w_s, the generated handwritten word image X_s should replicate the unique handwriting style of w_s while maintaining the content of $\mathcal{A}$, where the textual content $\mathcal{A}$=$\{a_i\}_{i=1}^{L}$ spans a length L and a_i is from a character set without any restrictions.

To address this task, we incorporate high-frequency information to enhance the extraction of writing styles. Existing methods [4, 8, 12, 13, 24, 39] typically use a vanilla CNN or transformer encoder to directly process style images, often resulting in undesired style extraction. In contrast, we introduce a novel One-shot diffusion mimicker (One-DM) by innovating Laplacian high-frequency extraction and a gating mechanism. Our One-DM can effectively capture style features from a single reference while suppressing background noise.

3.1 Overall Scheme

Our high-level idea focuses on incorporating high-frequency information from style reference images to enhance the extraction of style patterns. A straightforward implementation involves using a vanilla transformer encoder to extract style features from both the style image and its corresponding high-frequency image. This naive solution encounters two main issues: (1) the lack of efficient supervision objective still makes it challenging to accurately learn the writer's style patterns from the high-frequency image, and (2) the style features captured from the original image still retain undesirable noise backgrounds, which can adversely affect subsequent image generation performance.

To address the above-mentioned issues, we develop a more effective method, shown Figure 3. Our method consists of a style-enhanced module, a content encoder, a style-content fusion module, and a conditional diffusion module. Initially, we utilize the Laplacian kernel as a high-frequency filter to extract the high-frequency components from the style reference. Then, two parallel style encoders are employed to simultaneously extract the corresponding style features from both style reference and its high-frequency information. Since undesirable background noise is often present in style references, we design a gate mechanism to facilitate the transmission of informative style information while mitigating noise transmission. The style patterns in the high-frequency components are comparatively cleaner and more distinct, facilitating the observation of individual styles, such as character slant. For the observation, we propose a contrastive learning objective, termed LaplacianNCE $\mathcal{L}_{lapNCE}$, to enforce the more discriminative style learning from high-frequency components.

Regarding content guidance, we render the given string $\mathcal{A}$ into a unifont image, as done in VATr [39]. Briefly, Unifont's key advantage is that it covers all Unicode characters, allowing our method to transform any user inputs into corresponding images. We further feed render results into a content encoder that combines a ResNet18 [16] with a Transformer encoder. This process first involves using a ResNet18 to handle each character image in parallel, followed by concatenating them to form word sequence features. The Transformer encoder then processes these features to extract an informative content feature $E=\{e_i\}_{i=1}^{L} \in \mathbb{R}^{L\times c}$ with a global context, where c is the channel dimension. After obtaining style features and content guidance, we seamlessly fusion them using a style-content fusion module. The fused results then guide the denoising process of the conditional diffusion model to synthesize the desired handwritten text images. The denoising process is supervised by the reconstruction loss $\mathcal{L}_{rec}$.

To summarize, the overall training objective of our method combines all two loss functions:

$$\mathcal{L} = \mathcal{L}_{lapNCE} + \mathcal{L}_{rec}.$$

3.2 Style-enhanced Module

We propose the style-enhanced module to enhance style extraction by incorporating high-frequency components $H_s \in \mathbb{R}^{h\times w\times c}$ from a reference image $I_s \in \mathbb{R}^{h\times w\times c}$ since clearer style patterns are presented in high-frequency components like character slanting and shape connections (cf. Figure 2). As shown in Figure 3, we use a Laplacian kernel as a high-frequency filter to extract H_s from I_s. The Laplacian kernel excels in extracting high-frequency information without the need for Fast Fourier Transform (FFT) and parametric separation in the frequency domain. Then, two style encoders, $\mathcal{E}_{spa}$ and $\mathcal{E}_{fre}$, each a combination of CNN and transformer, process the I_s and H_s, respectively. This independent processing leads to the extraction of distinct style features:$F_{spa}=\{f_{spa}^{i}\}_{i=1}^{d} \in \mathbb{R}^{d\times c}$ from $\mathcal{E}_{spa}$ and $F_{fre}=\{f_{fre}^{i}\}_{i=1}^{d} \in \mathbb{R}^{d\times c}$ from $\mathcal{E}_{fre}$, where $d = h \times w$. While structurally identical, $\mathcal{E}_{spa}$ and $\mathcal{E}_{fre}$ do not share weights with each other. Then, the proposed $\mathcal{L}_{lapNCE}$ forces $\mathcal{E}_{fre}$ to focus on extracting discriminative style features

from H_s. A gate mechanism is designed to selectively filter out background noise from the reference style features, allowing only meaningful style patterns to pass.

Laplacian Contrastive Learning. The goal of the proposed $\mathcal{L}_{lapNCE}$ is to guide the high-frequency style encoder $\mathcal{E}_{fre}$ in learning more discriminative style features from the high-frequency information. Thus, we propose to bring closer extracted style features F_{fre} belonging to the same writer, while distancing those from different writers. We formulate our $\mathcal{L}_{lapNCE}$ as follows:

$$\mathcal{L}_{lapNCE}=-\frac{1}{N}\sum_{i\in M}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i\cdot z_a/\tau)}. \tag{1}$$

In detail, $i\in M=\{1...N\}$ is the index of any element in the mini-batch with size N and $A(i)=M\backslash\{i\}$ is other indices distinct from i. z_i is an anchor sample belonging to writer w_i and $P(i)=\{p\in A(i):w_p=w_i\}$ is its in-batch positive sample set and the other $A(i)\backslash P(i)$ is its negative set. Here, $z=Proj(F_{fre})$, where $Proj$ is a learnable multi-layer perceptron (MLP), τ is a scalar temperature parameter and the $\cdot$ symbol denotes the inner product.

Gate mechanism. As illustrated in Figure 2, the stroke areas of characters in reference images are typically sparse, with background noise interfering with the extraction of character style features. To address this challenge, we propose a gate mechanism to selectively filter the information of the sample I_s, as shown in Figure 3. Specifically, the extracted sample style features F_{spa} are fed into a gate layer, consisting of a learnable fully connected layer and followed by a sigmoid activation, to obtain the corresponding gate units $W=\{w_i\}_{i=1}^{d}\in\mathbb{R}^d$. Each unit w_i determines the pass rate for the corresponding f_{spa}^i, allowing for a higher pass rate where w_i is larger. This design effectively enables informative style features $\hat{F}_{spa}=\{\hat{f}_{spa}^i\}_{i=1}^{d}$ to be extracted while suppressing extraneous background noise, where $\hat{f}_{spa}^i=f_{spa}^i\cdot w_i$.

3.3 Style-content Fusion Module

Upon acquiring textual content feature E and two style features $\hat{F}_{spa}$ and F_{fre}, we integrate all features within two multi-head attention mechanisms to guide the denoising generation process of the diffusion model, as shown in Figure 3. Specifically, the first cross-attention module uses the textual content E as queries to identify the most relevant style information in the style reference, thereby inferring the style attributes corresponding to each character. For instance, if the textual content is 'a', it prioritizes searching for style features of characters like 'a', 'b', 'd', 'g' in the style reference, due to these characters appearing similar looped structures, implying more comparable style attributes. This process (cross-attention in Figure 3) is represented as:

$$O=Atten_1(Q_1=E,K_1=V_1=\hat{F}_{spa}+F_{fre}). \tag{2}$$

Subsequently, we obtain the initial fusion embedding between content and style guidance by simply summing O and E. The merged intermediate vector is then

employed as the query, key, and value in the self-attention mechanism to facilitate comprehensive interaction of information. Finally, the blended embedding g serves as the condition of the diffusion process. The second multi-head attention (self-attention in Figure 3) is defined as:

$$g = Atten_2(Q_2 = K_2 = V_2 = O + E). \tag{3}$$

3.4 Conditional Diffusion Model

The goal of the conditional diffusion model p_θ is to generate realistic images of handwritten text, guided by acquired conditions g. Specifically, as shown in Figure 3, under the guidance of g, p_θ executes a denoising generative process with T steps, starting from a sampled Gaussian noise x_T and progressively denoises to obtain the desired handwritten text x_0:

$$p_\theta(x_0|g) = \int p_\theta(x_{0:T}|g) d_{x_{1:T}}, \tag{4}$$

$$p_\theta(x_{0:T}|g) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, g), \tag{5}$$

$$p_\theta(x_{t-1}|x_t, g) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, g, t), \Sigma_\theta(x_t, g, t)). \tag{6}$$

The denoising process aims to learn how to reverse a predefined forward process, as described in DDPM [18]. The forward process is modeled as a fixed Markov chain, where noise conforming to a normal distribution is incrementally added to x_{t-1} to derive x_t. This can be mathematically expressed as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}), \tag{7}$$

where the noise is characterized by a variance schedule β_t. During training, a variational bound on the maximum likelihood objective is applied to guide the generation process to recover x_0 from the standard Gaussian noise x_T conditioned on g. We give the training objective as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{t,q} \left\| \mu_t(x_t, x_0) - \mu_\theta(x_t, g, t) \right\|_2^2, \tag{8}$$

where $\mu_t(x_t, x_0)$ is the mean of the forward process posterior $q(x_t|x_{t-1})$, which has a closed-form solution [18]. We put more details in supplementary material.

4 Experiments

4.1 Experimental Settings

Dataset. To evaluate our One-DM in handwritten text generation, we use the widely-used handwriting dataset IAM [36] and CVL [27]. IAM consists of 62,857 English word images from 500 unique writers. Following [4, 24, 39], we use words

from 339 writers for training and the remaining 161 writers for testing. CVL includes texts by 310 writers in English and German. We use the English portion, totaling 84,514 words, and follow CVL's standard split, with 283 writers for training and 27 for testing. Throughout all experiments, we resize images to 64 pixels in height, preserving their aspect ratio.

Evaluation metrics. We use the Fréchet Inception Distance (FID) [17] and Geometry Score (GS) [26] to assess the generation quality, following the settings of [4, 24, 39]. We also conduct user studies to quantify the subjective quality of the generated handwritten text images in supplementary.

Implementation details. In our all experiments, we only use single style reference sample. We first train our model for 700 epochs (batch size of 384) with classifier-free guidance strategy [19] under the guidance scale of 0.25, and then fine-tune the model for 4500 iterations (batch size of 128) using a text recognizer [41] with a CTC loss [15, 21, 58], on four RTX3090 GPUs. The fine-tuning process forces our One-DM to generate readable text with accurate content. The optimizer is AdamW [33], with a learning rate of 10^{-4}. During inference, each style sample is randomly sampled from the target-writer. To speed up the sampling, we use the denoising diffusion implicit model (DDIM) [43] with 50 steps. More details are provided in supplementary material.

Compared methods. We compare our One-DM with state-of-the-art handwritten text generation methods, including GAN-based methods (i.e., GANwriting [24], HWT [4], VATr [39], TS-GAN [8], HiGAN+ [13]), and diffusion-based methods (*i.e.* GC-DDPM [10] and WordStylist [37]). For a fair comparison, we configure all methods to generate images with a height of 64 pixels in supplementary. Additionally, in supplementary, we compare a variant of One-DM that generates 32-pixel high images with official VATr [39] and HWT [4].

4.2 Main Results

Styled Handwritten Text Generation. Initially, we assess our One-DM for producing styled handwritten text images, aiming to replicate both the style and content in the generated images. Following [4, 24, 39], we first calculate FID between generated and real samples for each writer separately and finally average them. In line with the previous works [4, 24, 39], our experiments on the IAM dataset are divided into four distinct scenarios: IV-S, IV-U, OOV-S, OOV-U. Among these four scenarios, OOV-U represents the most challenging case where both the target style and the words are entirely unseen during training. For the CVL dataset, we directly report the results of all methods on the test set.

We first report the quantitative results on the IAM dataset in Table 1. We can observe that our One-DM outperforms all the competitors in all settings. Notably, it significantly exceeds one-shot methods in all scenarios. Impressively, our One-DM also holds a substantial advantage over few-shot methods (GANwriting [24], HWT [4], VATr [39]), which use 15-x more reference samples for style guidance, in IV-S and OOV-S settings. Even in the most challenging OOV-U scenario, our One-DM leads the second-best, VATr, by a large margin (102.75 *vs.* 108.76), demonstrating the superior performance of our One-DM in stylized

Table 1: Comparisions with state-of-the-art methods on styled and style-agnostic handwritten text generation in the IAM dataset. Note that, all methods are trained on the same training set used in GANwriting [24], HWT [4], and VATr [39].

Method	Shot	Styled Evaluation				Style-agnostic	
		IV-S	IV-U	OOV-S	OOV-U	FID↓	GS↓
TS-GAN [8]	One	118.56	128.75	127.11	136.67	20.65	4.88×10^{-2}
GANwriting [24]	Few	120.07	124.30	125.87	130.68	28.37	5.67×10^{-2}
HiGAN+ [13]	One	117.33	116.95	121.55	121.48	22.95	2.06×10^{-2}
GC-DDPM [10]	One	99.86	105.73	112.52	118.39	19.05	1.31×10^{-2}
WordStylist [37]	One	98.10	104.27	109.45	115.52	18.58	2.85×10^{-2}
HWT [4]	Few	109.25	106.90	116.55	113.52	18.99	4.41×10^{-3}
VATr [39]	Few	103.75	101.73	111.64	108.76	16.03	1.74×10^{-2}
Ours (One-DM)	One	**89.47**	**98.36**	**93.30**	**102.75**	**15.73**	$\mathbf{1.98\times10^{-3}}$

handwritten text generation. Similarly, our method outperforms HWT and VATr on the CVL dataset, achieving the lowest FID score, as shown in Table 4.

We provide qualitative results to intuitively explain the benefit of our One-DM, in Figure 4. GANwriting struggles to capture the style patterns of reference samples, such as character slant and occasionally produces unclear character shapes. HiGAN+ more consistently generates characters with correct content, but the character spacing within the generated words lacks realism. WordStylist typically produces images with noticeable background noise. HWT and VATr can produce satisfactory handwritten words in terms of content accuracy and style mimicry; however, their downside is the tendency to create smoother character appearances. Compared to HWT and VATr, our synthesized samples excel in more authentic character ink color and stroke thickness. However, some generated samples by our One-DM are visibly different in ink color. We provide comprehensive explanations in supplementary. We further display more qualitative comparisons between our method and few-shot methods in Figure 5.

Style-agnostic Handwritten Text Generation. We further evaluate our One-DM to generate realistic handwritten text images, irrespective of style imitation. For this purpose, we compute the FID and GS on the IAM test set, under the same conditions as applied to ScrabbleGAN [11] (FID: 20.72, GS:2.56×10^{-2}), capable of generating handwritten texts with randomly sampled styles. Specifically, each method generates 25k random samples to calculate FID against 25k samples from the test set, and 5k random samples for GS calculation in comparison with 5k test set samples. As presented in Table 1, our One-DM achieves the best results in both FID and GS metrics, further demonstrating its capability to generate higher-quality handwritten text images.

4.3 Analysis

In this section, we conduct ablation studies to analyze our One-DM. More analyses are provided in suplementary, including generalization evaluation on different style backgrounds, generation quality assessment through OCR performance, failure case analysis, and the effects of different designs (*e.g.*, high-frequency filter, style-content fusion mechanism, and style input sample length).

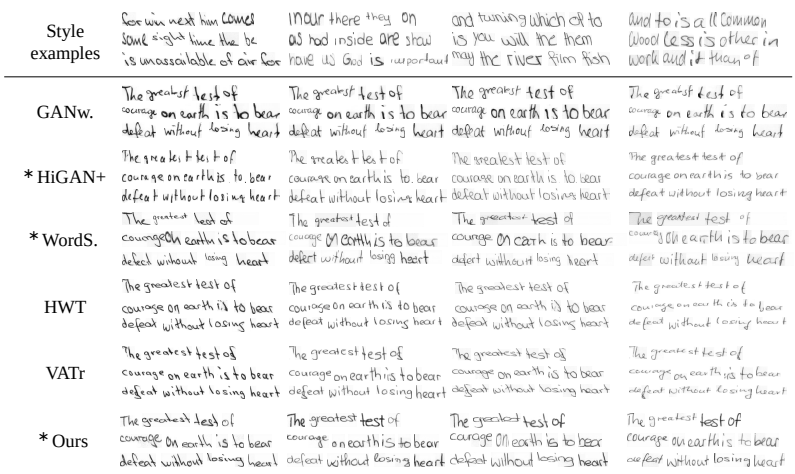


Fig. 4: Qualitative comparisons between our method with state-of-the-art methods on handwritten text generation with both specific textual content and desired handwriting style in the IAM dataset. We utilize the identical guiding text, '*The greatest test of courage on earth is to bear defeat without losing heart*,' across all handwriting generation methods, directing them to produce text in varied styles. Better zoom in 200%. $*$ denotes one-shot methods, while others are few-shot methods.

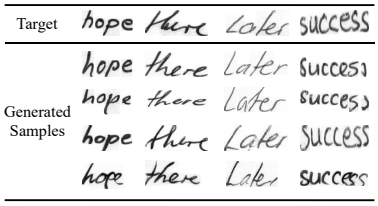


Fig. 5: Each row shows results from our One-DM and few-shot methods on IAM dataset; readers are invited to identify our method. The answer is at the paper's end.

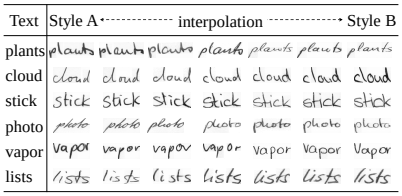


Fig. 6: We provide style interpolation results generated by our One-DM. Different individual writing styles are extracted from the IAM test dataset.

Quantitative evaluation of Laplacian branch and gate mechanism. We conduct various ablation experiments on the IAM dataset to evaluate the effect of distinct components within our approach. We provide the quantitative result in Table 2. We find that: (1) The inclusion of both Laplacian branch and gate mechanism enhances the quality of the generated handwritten text images, improving FID by 3.92 and 2.71, respectively. (2) Integrating the Laplacian branch with the gate mechanism further boosts the generative performance.

Qualitative evaluation of Laplacian branch and gate mechanism. To further analyze each module in our One-DM, we conduct visual ablation experiments. As shown in Table 2. we can observe that firstly, after adding a gate mechanism, background noise can be somewhat suppressed, resulting in characters with relatively clean backgrounds. Then, the independent addition of the Laplacian branch helps the model learn cursive connections and other

Table 2: Ablation study. Effect of the Laplacian branch and gate mechanism on the IAM dataset under the OOV-U setting, as done in [4].In the middle, we showcase the generated samples of each component.

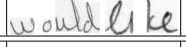
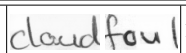
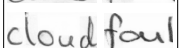
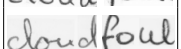
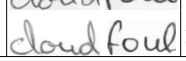
Base	$\mathcal{E}_{fre}$	Gate	Style samples	FID ↓
				
✓				108.44
✓		✓		105.73
✓	✓			104.52
✓	✓	✓		102.75

Table 3: Effect of each part of our Laplacian branch on IAM dataset under OOV-U setting.

H_s	$\mathcal{L}_{lapNCE}$	FID ↓
		108.44
✓		106.16
	✓	107.58
✓	✓	104.52

Table 4: Comparisons with competitors on the CVL dataset.

Method	FID ↓
HWT [4]	58.22
VATr [39]	54.44
Ours (One-DM)	51.78

style patterns. Finally, our method integrated the Laplacian branch and gate mechanism, which can generate the highest quality handwritten text images.

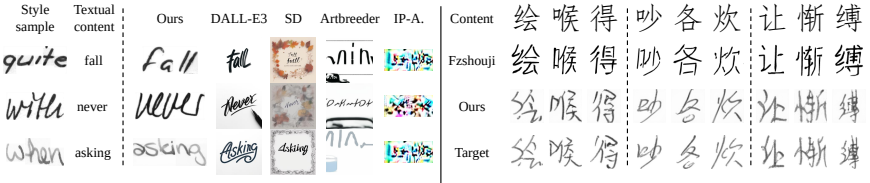
Discussions of the Laplacian branch. Our Laplacican branch consists of two key components: utilizing high-frequency images H_s and the Laplacian contrastive learning loss $\mathcal{L}_{lapNCE}$. In the previous ablation study, they are always combined. We further conduct exploration experiments to explain why they cannot be separated. As reported in Table 3, combining H_s and $\mathcal{L}_{lapNCE}$ maximizes effectiveness; separating them significantly reduces performance. Without the guidance of $\mathcal{L}_{lapNCE}$, extracting discriminative features from H_s is challenging. Likewise, directly applying $\mathcal{L}_{lapNCE}$ on original images leads to an undesired style feature extraction, as original images have less clear style patterns than H_s.

Discussions about learning style from a single reference. We are quite surprised that One-DM, with just a single reference sample, even exceeds the generative performance of few-shot methods. We provide the potential reason analysis below. Firstly, One-DM learns a meaningful style latent space, wherein new styles can be generated based on seen styles (cf. Figure 6). Then, through our style-enhanced module, One-DM effectively extracts the style feature from a single example and maps it to a position close to the exemplary writer in the feature space, thereby producing high-quality styled handwritten text images.

4.4 Comparisons with SOTA Industrial Methods

To highlight the superiority of our method, we compare One-DM with leading industrial image generation methods that are trained on tremendously large datasets (including numerous text-centric images), including two prominent text-to-image methods, DALL-E3 [3] and Stable Diffusion [42] (SD), and two popular style transfer methods, Artbreeder[1] and IP-Adapter[2] (IP-A.), on the IAM dataset. Further experiment details are in supplementary material.

[1] https://www.artbreeder.com/
[2] https://github.com/tencent-ailab/IP-Adapter



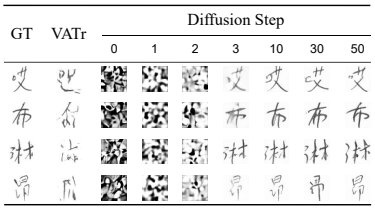
(a) Comparisons on handwritten text generation (b) Comparisons on Chinese handwriting generation

Fig. 7: (a) Qualitative comparisons between our method with SOTA industrial image generation methods, including DALL-E3 [3], Stable Diffusion [42], Artbreeder[1] and IP-Adapter[2] on handwritten text generation with both controllable styles and contents. (b) Qualitative comparisons with Fzshouji[3] on ICDAR- 2013 competition database [52].

Table 5: Quantitative comparisons with competitors on styled handwritten character generation in terms of FID.

Method	Chinese	Japanese
GANwriting [24]	116.49	111.86
HWT [4]	165.74	148.66
VATr [39]	139.91	124.98
WordStylist [37]	34.61	101.93
Ours (One-DM)	27.24	95.43

Table 6: Qualitative comparisons with VATr [39] on Chinese dataset.



As shown in Figure 7(a), our method excels industrial methods in style mimicry and content preservation. The performance of IP-A. is the poorest, often producing distorted images. Artbreeder can replicate the color of the strokes from style samples, but fails in content preservation. DALL-E3 and SD generate characters with accurate content, but often mismatch style details with references, such as character spacing and stroke width, with SD often generating extra backgrounds. Besides, we compare Fzshouji[3], an advanced industrial method designed for Chinese handwriting generation. As shown in Figure 7(b), our method outperforms Fzshouji in replicating character details and ink color.

4.5 Applications to Other Languages

In this section, we evaluate whether One-DM can be used to generate languages other than English. We further conduct experiments in the Chinese (*i.e.* ICDAR-2013 competition database [52]) and Japanese(*i.e.* UP_Kuchibue database [22]) datasets. We use FID to assess generated samples of each writer, then average them. More experiment details can be seen in supplementary material.

For the Chinese handwritten character generation task, as shown in Table 5, we find that our One-DM outperforms the second-best method by a large margin, achieving a 7.37 lower FID. From Figure 8 (a), we can observe that our One-DM generates characters that closely match the target images in geometric shape and character slant. In contrast, handwritings from HWT and VATr exhibit noticeable artifacts like blurring and collapsed structures. GANwriting tends to

[3] https://www.fzshouji.com/make

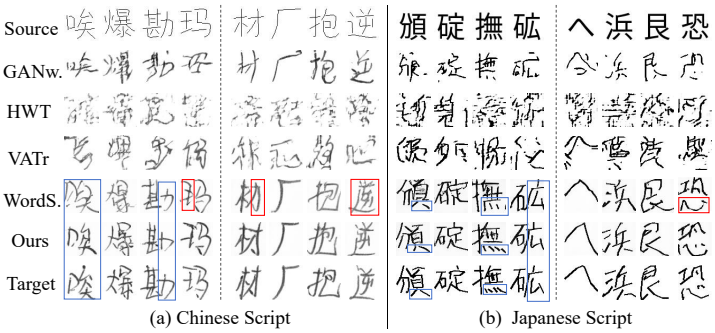


Fig. 8: Comparisons with GANwriting [24], HWT [4], VATr [39], and WordStylist [37] for styled handwritten character generation on Chinese and Japanese scripts. The red boxes highlight failures of structure preservation, while blue boxes highlight comparisons between the style patterns of targets and generated characters.

miss strokes in its handwritings. WordStylist sometimes struggles to accurately mimic style patterns and tends to generate characters with incorrect radicals. Table 5 and Figure 8 (b) further verify the effectiveness of One-DM for Japanese handwriting generation. We also achieve the lowest FID score, and our generated Japanese samples excel in both content preservation and style imitation.

We further investigate why diffusion-based methods (Our One-DM and WordStylist) that require only one single sample outperform few-shot GAN-based methods (i.e., GANwriting, HWT, and VATr) so significantly in generating Chinese and Japanese characters. The lower performance of GAN-based methods on Chinese and Japanese characters may stem from their vanilla convolutional architectures struggling with Chinese and Japanese characters' complex geometry, as noted in [49]. In contrast, our One-DM breaks down the generation of Chinese and Japanese characters into simpler steps. For instance, as demonstrated in Table 6, during the early stages of the diffusion generation process, the model first attempts to generate a rough Chinese handwritten character. It then continues to refine the writing style (e.g., character shape and stroke color) under the condition guidance, until it synthesizes handwritings that are satisfactory.

5 Conclusion

In this paper, we introduce a novel One-DM for handwritten text generation, requiring only one style reference to produce realistic handwritten text images[4]. We enhance style extraction by incorporating high-frequency components from the style reference. For high-frequency components with distinct style patterns, we employ laplacian contrastive learning to capture more discriminative style features. Moreover, a gate mechanism improves the transfer of informative features from the reference, reducing background noise. Our One-DM outperforms few-shot methods in multiple language scripts. In the future, we aim to explore the potential of One-DM in font generation and vector font creation tasks.

[4] In Figure 5, the second from bottom row belongs to our method.

Acknowledgments The research is partially supported by National Key Research and Development Program of China (2023YFC3502900), National Natural Science Foundation of China (No.62176093, 61673182), Key Realm Research and Development Program of Guangzhou (No.202206030001), Guangdong-Hong Kong-Macao Joint Innovation Project (No.2023A0505030016).

References

1. Aksan, E., Pece, F., Hilliges, O.: Deepwriting: Making digital ink editable via deep generative modeling. In: ACM Conf. Human Factors Comput. Sys. pp. 1–14 (2018)
2. Alonso, E., Moysset, B., Messina, R.: Adversarial generation of handwritten text images conditioned on sequences. In: ICDAR. pp. 481–486 (2019)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. **2**(3), 8 (2023)
4. Bhunia, A.K., Khan, S., Cholakkal, H., Anwer, R.M., Khan, F.S., Shah, M.: Handwriting transformers. In: ICCV. pp. 1086–1094 (2021)
5. Cao, J., Mo, L., Zhang, Y., Jia, K., Shen, C., Tan, M.: Multi-marginal wasserstein gan. In: NeurIPS. vol. 32 (2019)
6. Chen, Z., Yang, D., Liang, J., Liu, X., Wang, Y., Peng, Z., Huang, S.: Complex handwriting trajectory recovery: Evaluation metrics and algorithm. In: ACCV. pp. 1060–1076 (2022)
7. Dai, G., Zhang, Y., Wang, Q., Du, Q., Yu, Z., Liu, Z., Huang, S.: Disentangling writer and character styles for handwriting generation. In: CVPR. pp. 5977–5986 (2023)
8. Davis, B.L., Morse, B.S., Price, B.L., Tensmeyer, C., Wigington, C., Jain, R.: Text and style conditioned GAN for the generation of offline-handwriting lines. In: BMVC (2020)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
10. Ding, H., Luan, B., Gui, D., Chen, K., Huo, Q.: Improving handwritten ocr with training samples generated by glyph conditional denoising diffusion probabilistic model. In: ICDAR. pp. 20–37 (2023)
11. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: CVPR. pp. 4324–4333 (2020)
12. Gan, J., Wang, W.: Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles. In: AAAI. pp. 7484–7492 (2021)
13. Gan, J., Wang, W., Leng, J., Gao, X.: Higan+: Handwriting imitation gan with disentangled representations. ACM TOG **42**(1), 1–17 (2022)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
15. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: ICML. pp. 369–376 (2006)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

17. 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: Advances in neural information processing systems (2017)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: arXiv (2022)
20. Huang, H., Yang, D., Dai, G., Han, Z., Wang, Y., Lam, K.M., Yang, F., Huang, S., Liu, Y., He, M.: Agtgan: Unpaired image translation for photographic ancient character generation. In: ACM MM. pp. 5456–5467 (2022)
21. Huang, S., Luo, Y., Zhuang, Z., Yu, J.G., He, M., Wang, Y.: Context-aware selective label smoothing for calibrating sequence recognition model. In: ACM MM (2021)
22. Jaeger, S., Nakagawa, M.: Two on-line japanese character databases in unipen format. In: ICDAR. pp. 566–570 (2001)
23. Kang, L., Riba, P., Rusinol, M., Fornes, A., Villegas, M.: Content and style aware generation of text-line images for handwriting recognition. IEEE TPAMI **44**(12), 8846–8860 (2021)
24. Kang, L., Riba, P., Wang, Y., Rusinol, M., Fornés, A., Villegas, M.: Ganwriting: content-conditioned generation of styled handwritten word images. In: ECCV. pp. 273–289 (2020)
25. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: NeurIPS. pp. 18661–18673 (2020)
26. Khrulkov, V., Oseledets, I.: Geometry score: A method for comparing generative adversarial networks. In: International conference on machine learning. pp. 2621–2629. PMLR (2018)
27. Kleber, F., Fiel, S., Diem, M., Sablatnig, R.: Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In: ICDAR. pp. 560–564 (2013)
28. Kotani, A., Tellex, S., Tompkin, J.: Generating handwriting via decoupled style descriptors. In: ECCV. pp. 764–780 (2020)
29. Li, D., Chen, G., Wu, X., Yu, Z., Tan, M.: Face anti-spoofing with cross-stage relation enhancement and spoof material perception. Neural Networks **175**, 106275 (2024)
30. Lin, T., Ma, Z., Li, F., He, D., Li, X., Ding, E., Wang, N., Li, J., Gao, X.: Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In: CVPR. pp. 5141–5150 (2021)
31. Liu, Z., Chen, K., Zhang, Y., Han, J., Hong, L., Xu, H., Li, Z., Yeung, D.Y., Kwok, J.: Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. arXiv (2023)
32. Liu, Z., Jia, W., Yang, M., Luo, P., Guo, Y., Tan, M.: Deep view synthesis via self-consistent generative network. IEEE TMM **24**, 451–465 (2021)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: arXiv (2017)
34. Luhman, T., Luhman, E.: Diffusion models for handwriting generation. arXiv (2020)
35. Luo, C., Zhu, Y., Jin, L., Li, Z., Peng, D.: Slogan: handwriting style synthesis for arbitrary-length and out-of-vocabulary text. IEEE Trans. Neural Networks Learn. Syst. (2022)
36. Marti, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002)

37. Nikolaidou, K., Retsinas, G., Christlein, V., Seuret, M., Sfikas, G., Smith, E.B., Mokayed, H., Liwicki, M.: Wordstylist: Styled verbatim handwritten text generation with latent diffusion models. In: ICDAR. pp. 384–401 (2023)
38. Pan, Z., Ji, Z., Liu, X., Bai, J., Liu, C.L.: Visa: Visual and semantic alignment for robust scene text recognition. In: ICDAR. pp. 223–242 (2023)
39. Pippi, V., Cascianelli, S., Cucchiara, R.: Handwritten text generation from visual archetypes. In: CVPR. pp. 22458–22467 (2023)
40. Ren, M.S., Zhang, Y.M., Wang, Q.F., Yin, F., Liu, C.L.: Diff-writer: A diffusion model-based stylized online handwritten chinese character generator. In: International Conference on Neural Information Processing. pp. 86–100 (2023)
41. Retsinas, G., Sfikas, G., Gatos, B., Nikou, C.: Best practices for a handwritten text recognition system. In: International Workshop on Document Analysis Systems. pp. 247–259 (2022)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
44. Tang, S., Lian, Z.: Write like you: Synthesizing your cursive online chinese handwriting via metric-based meta learning. In: Comput. Graph. Forum. pp. 141–151 (2021)
45. Tang, S., Xia, Z., Lian, Z., Tang, Y., Xiao, J.: Fontrnn: Generating large-scale chinese fonts via recurrent neural network. In: Comput. Graph. Forum. pp. 567–577 (2019)
46. Tolosana, R., Delgado-Santos, P., Perez-Uribe, A., Vera-Rodriguez, R., Fierrez, J., Morales, A.: Deepwritesyn: On-line handwriting synthesis via deep short-term representations. In: AAAI. pp. 600–608 (2021)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
48. Xie, X., Fu, L., Zhang, Z., Wang, Z., Bai, X.: Toward understanding wordart: Corner-guided transformer for scene text recognition. In: ECCV. pp. 5456–5467 (2022)
49. Xie, Y., Chen, X., Sun, L., Lu, Y.: Dg-font: Deformable generative networks for unsupervised font generation. In: CVPR (2021)
50. Yang, Y., Liu, D., Zhang, S., Deng, Z., Huang, Z., Tan, M.: Hilo: Detailed and robust 3d clothed human reconstruction with high-and low-frequency information of parametric models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10671–10681 (2024)
51. Yang, Y., Zhang, S., Huang, Z., Zhang, Y., Tan, M.: Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15901–15911 (2023)
52. Yin, F., Wang, Q.F., Zhang, X.Y., Liu, C.L.: Icdar 2013 chinese handwriting recognition competition. In: International Conference on Document Analysis and Recognition. pp. 1464–1470 (2013)
53. Zhang, Y., Hooi, B.: Hipa: Enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. arXiv (2023)
54. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In: NeurIPS. pp. 34077–34090 (2022)
55. Zhang, Y., Hooi, B., Hu, D., Liang, J., Feng, J.: Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In: NeurIPS. pp. 29848–29860 (2021)

56. Zhao, B., Tao, J., Yang, M., Tian, Z., Fan, C., Bai, Y.: Deep imitator: Handwriting calligraphy imitation via deep attention networks. Pattern Recognition **104**, 107080 (2020)
57. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: CVPR. pp. 14235–14245 (2023)
58. Zhuang, Z., Liu, Z., Lam, K.M., Huang, S., Dai, G.: A new semi-automatic annotation model via semantic boundary estimation for scene text detection. In: ICDAR. pp. 257–273 (2021)