Hierarchical Style Aggregation for Versatile Chinese Handwriting Generation

Jiangpeng Wang, Fei Gao*, and Nannan Wang*

Xidian University, Xi'an 710126, China
jiangpengwang@stu.xidian.edu.cn, {fgao, nnwang}@xidian.edu.cn

Abstract. Few-shot Chinese handwriting generation aims to render characters in a target style using limited references. Existing image-level methods often suffer from limited stylistic fidelity and poor generalization. We propose Hierarchical Style Aggregation for Chinese handwriting generation (HiSAC), a versatile framework that decouples style modeling from task-specific generation. Our core idea is to leverage the compositional nature of Chinese characters: by decomposing characters into radicals, components, and strokes, we establish fine-grained spatial correspondences between target content and reference glyphs at each level, enabling precise style transfer. To further capture nuanced style characteristics, we design a multi-band frequency encoder that extracts stylistic representations across different spectral ranges. The aggregated multi-granularity features can be seamlessly integrated with task-specific decoders (e.g., Transformer for online trajectory generation, Diffusion for offline image synthesis) without architectural redesign. By jointly modeling style in both spatial and frequency domains, our approach enhances realism in global structure and stroke dynamics. Extensive experiments demonstrate that HiSAC outperforms existing methods both quantitatively and qualitatively, and generalizes well to out-of-vocabulary and cross-language scripts. Code has been released at `https://github.com/IIP-Lab-XDU/HiSAC`.

Keywords: Chinese Handwriting Generation · Few-shot Learning · Hierarchical Style Aggregation · Versatile Framework

1 Introduction

A fundamental challenge in document analysis and understanding is the recognition of handwritten text under long-tailed and open-set conditions [28]. Few-shot handwritten character generation, which aims to synthesize a writer's style from only a few samples, presents a viable pathway to address this issue. However, the vast vocabulary and complex hierarchical composition of Chinese characters (radicals, components, strokes) make this task especially challenging [41]. Existing approaches fall into two categories. *Offline* methods, typically based

* Corresponding authors.

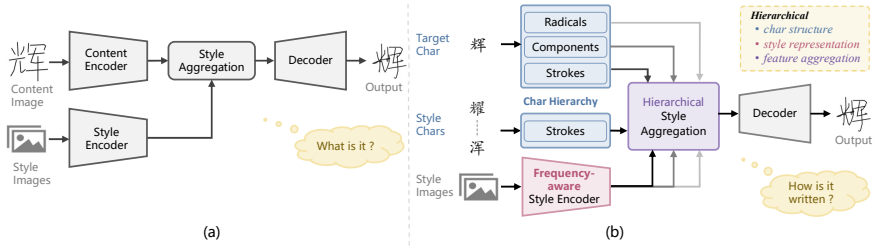

Fig. 1: Architecture differences between (a) previous pipeline and (b) ours. Previous methods use a content encoder trained for character classification, focusing on *what it is*. Our model decomposes characters to learn *how it is written*.

on Generative Adversarial Networks (GANs) [2, 4, 13, 24, 31] or diffusion models [7, 8, 11, 18, 33, 34, 60], often struggle to capture the compositional nature of Chinese glyphs. In contrast, *online* methods [9, 22, 30, 48, 58, 59] simulate the writing process using RNNs or Transformers, effectively modeling temporal dynamics but requiring trajectory data and being limited to stroke-based synthesis.

Most existing methods encode content using networks pre-trained for character recognition [7, 9, 48], focusing on *what the character is* rather than *how it is structured* (Fig. 1), while relying on global image-level features to represent style [33, 34]. This design disregards essential structural priors and fails to capture local stylistic details. As illustrated in Fig. 2, generating a stylized Chinese character involves complex correspondences between the target content and reference glyphs across multiple structural scales—from overall layout to fine-grained stroke curvature. In few-shot settings, this mismatch between existing methods and the intrinsic multi-scale nature of handwriting leads to inaccurate stylization and poor generalization to unseen characters [4, 39]. Although recent efforts have begun to incorporate structural information [30, 52], they still fall short of explicitly modeling compositional and stylistic hierarchies in an end-to-end manner.

To address these limitations, we propose *Hierarchical Style Aggregation for Chinese handwriting generation* (HiSAC). Inspired by the compositional nature of Chinese characters, HiSAC decomposes each character into three hierarchical levels, i.e., radicals, components, and strokes (Fig. 2), to capture layout configuration, structural composition, and writing curvature. Both target and reference characters are represented as structured sequences at each level, and the core *Hierarchical Style Aggregation* (HiSA) module leverages cross-scale correspondences between these sequences to transfer style information precisely and coherently. Beyond spatial structure, writing style itself manifests across multiple scales, from global posture to local stroke dynamics. We therefore design a multi-band *Frequency-Aware Style Encoding* (FASE) that extracts stylistic features across different frequency ranges. By jointly modeling style in both spatial and frequency domains, HiSAC achieves accurate and fine-grained style transfer. Finally, the aggregated multi-granularity features can be seamlessly integrated with task-specific decoders, such as a Transformer for online trajectory gener-

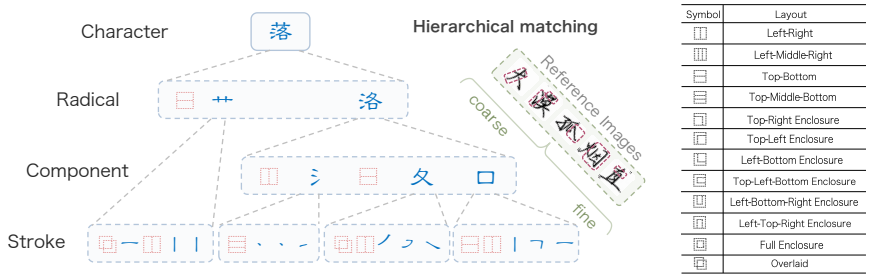

Fig. 2: Hierarchical char representation of a Chinese character and cross-scale correspondences between the target character and the reference glyphs.

ation [50] or a Diffusion model for offline image synthesis [21], without architectural redesign. Extensive experiments demonstrate that HiSAC consistently outperforms existing methods in both quantitative metrics and visual quality, while generalizing robustly to out-of-vocabulary and cross-language scripts.

Our key contributions are summarized as follows:

- We present a novel *Hierarchical Style Aggregation* (HiSA) mechanism. It achieves fine-grained and coherent style imitation by explicitly transferring and fusing style cues in a coarse-to-fine manner, guided by the compositional structure of Chinese characters.
- We propose a *multi-band Frequency-aware Style Encoder* (FASE) to learn hierarchical style representations, capturing the full spectrum of stylistic variations from global posture to local stroke dynamics.
- Our framework, HiSAC, unifies online and offline generation with a composition based design, achieving state-of-the-art performance and strong generalization to unseen and cross-language scripts.

2 Related Works

Offline Handwriting Generation. Early methods [13] focused on expressing textual content, either by synthesizing word-level images or assembling characters sequentially into words. Later, models like TSGAN [10] synthesize full lines of handwritten text by predicting blank spaces between characters, while GANwriting [24] breaks free from predefined vocabularies and enables generation of arbitrary words. SmartPatch [31] further reduces image artifacts with a lightweight discriminator. Subsequently, HiGAN series [15, 16] allow arbitrary-length text generation from a single reference image. Transformer-based models such as HWT [4] and VATr [39] improve text sequence modeling capabilities. In recent years, diffusion-based models [11, 33, 60] bring significant quality improvements by conditioning on style categories. The latest approaches, including One-DM [7] and DiffusionPen [34], eliminate reliance on fixed style IDs, enabling flexible generation in arbitrary writing styles.

Online Handwriting Generation. Early approaches based on RNNs, such as DeepWriting [1] and Drawing [58], model pen state dynamics for smooth

handwriting, but cannot support arbitrary style. FontRNN [49] introduces style transfer capabilities but lacks flexible style control. DeepImitator [59] employs CNN to extract style features encoded into a single vector, but it fails to capture fine-grained local details. WriteLikeYou [48] improves style discrimination using large-margin softmax and angular center losses. Recently, Transformer-based methods like SDT [9] adopt style disentanglement techniques to separately model writer style and character style. ElegantlyWritten [30] uses vector quantization in a Cartesian product space to compress style features while retaining essential information. DNA [22] incorporates component-level information of Chinese characters exclusively as an auxiliary signal for content representation.

Font Generation. Font generation methods, similar to handwriting generation, commonly disentangle content and style representations, as explored in previous works [14, 35, 51, 54, 55]. Some approaches [26, 37, 38, 47] further incorporate component-level structural information to enhance style modeling; however, they still depend on content image supervision or pre-generated content representations, which limits their flexibility. Although font and handwriting generation share similar objectives, handwriting generation poses greater challenges due to its higher stylistic variability and stronger individual differences.

3 Method

3.1 Overview

Given a set of reference glyph images $\mathcal{I}^{ref} = \{I_i^{ref}\}_{i=1}^{k}$ from a specific writer and a target character c^{tgt}, our goal is to generate a handwritten glyph for c^{tgt} that preserves its semantic content while mimicking the writer's unique style. To achieve this, we also leverage the reference characters $\mathcal{C}^{ref} = \{c_i^{ref}\}_{i=1}^{k}$ (*e.g.*, the UTF-8 codes corresponding to $\mathcal{I}^{ref}$) to establish structural matching between the target and reference characters.

As illustrated in Fig. 3, our framework first decomposes each character into three hierarchical levels (i.e., radicals, components, and strokes) and encodes them into structural embeddings. Currently, a *Frequency-Aware Style Encoder* (FASE) extracts multi-level style representations from reference glyph images $\mathcal{I}^{ref}$. A *Hierarchical Style Aggregation* (HiSA) module then fuses the structural and style features by performing cross-scale matching. Finally, the fused representation is passed to a decoder (based on a Transformer or Diffusion architecture) to synthesize the output, supporting both online and offline generation.

3.2 Hierarchical Char Representation

Our method begins with a hierarchical decomposition of Chinese characters, utilizing a fixed decomposition table[1] and a recursive algorithm. Figures 2 and 3 illustrate a standard three-layer case. For the **target character**, this decomposition yields three sequences representing structure at different granularities: a

[1] `https://babelstone.co.uk/CJK/IDS.TXT`

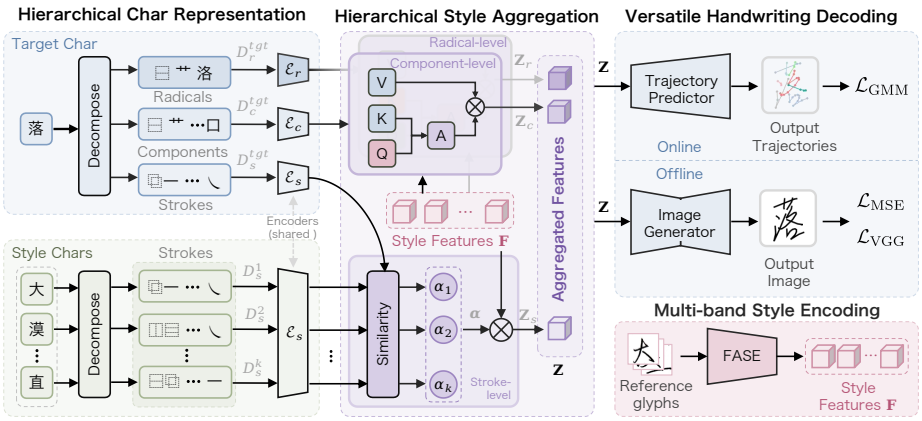

Fig. 3: The pipeline of the HiSAC framework. HiSAC unifies structure decomposition and *frequency-aware multi-band style encoding* (FASE) through the *hierarchical style aggregation* (HiSA) module to generate handwriting in both online and offline modes.

radical-level sequence $\mathcal{R}^{tgt} = \{r_j\}_{j=1}^{n_r}$ capturing global structure, a *component-level* sequence $\mathcal{C}^{tgt} = \{c_j\}_{j=1}^{n_c}$ encoding spatial layout, and a *stroke-level* sequence $\mathcal{S}^{tgt} = \{s_j\}_{j=1}^{n_s}$ detailing fine-grained elements. Correspondingly, each **reference character** is decomposed into a *stroke-level* sequence $\mathcal{S}_i^{ref} = \{s_{i,j}^{ref}\}_{j=1}^{m_i}$. Here, n_r, n_c, n_s, and m_i denote the respective sequence lengths, and the set of all reference strokes is $\mathcal{S}^{ref} = \{\mathcal{S}_i^{ref}\}_{i=1}^{k}$. Additionally, variable-length sequences at each level are handled via zero-padding to allow attention-based aggregation.

Notes: It is worth noting that both existing methods and our approach require the identities of target and reference characters (i.e., knowing which Chinese characters they are). However, existing methods fail to fully leverage this information, whereas we employ an automatic lookup table with *nearly zero computational overhead* to obtain hierarchical structural representations for each character, which are then used for cross-scale style transfer.

3.3 Frequency-Aware Multi-band Style Encoding (FASE)

Accurately capturing a writer's global and local stylistic characteristics is a primary challenge in handwriting generation. While recent works recognize the importance of frequency-domain analysis, they often treat it as an external input [7] or a post-processing step [5, 32, 53]. In contrast, rather than constructing a strict, non-overlapping, handcrafted frequency decomposition, we design a frequency-aware encoder, i.e., FASE, that intrinsically learns multi-band style representations. The pipeline of FASE is illustrated in Fig. 4 and detailed below.

Channel-Splitting. First, inspired by Inception architectures [45,46], we adopt a multi-branch design to capture complementary style cues at varying spatial scales. Specifically, the patch embeddings $\mathbf{X} \in \mathbb{R}^{n \times d}$ of a sample image are split along the channel dimension into three parts: $\mathbf{X}_c$, $\mathbf{X}_m$, and $\mathbf{X}_f$, with corresponding channel sizes d_c, d_m, and d_f, respectively, where $d = d_c + d_m + d_f$,

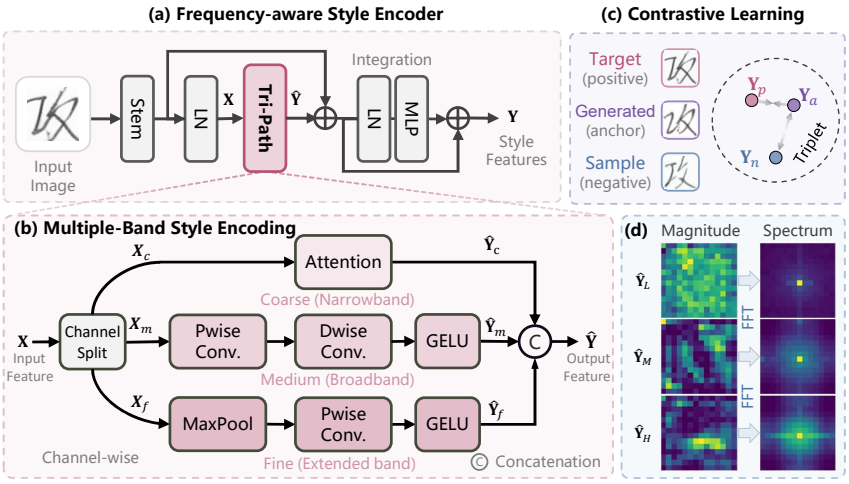

Fig. 4: Frequency-aware multi-band style encoding (FASE).

and n denotes the number of patch tokens. As illustrated in Fig. 4, these feature groups are fed into the coarse, medium, and fine branches, respectively, to extract multi-band style representations.

Coarse (Narrowband) Branch. The coarse branch design is particularly motivated by the spectral characteristics of *Vision Transformers*, which perform the attention mechanism across all channels and therefore tend to act as low-pass filters [12, 36]. The coarse branch thus captures overall writer characteristics, such as aspect ratio (*e.g.*, compact or stretched) and glyph slant (*e.g.*, upright or oblique). The attention output is computed as:

$$\hat{\mathbf{Y}}_c = \text{softmax}\left(\frac{\mathbf{Q}_c\mathbf{K}_c^\top}{\sqrt{d_c}}\right)\mathbf{V}_c, \tag{1}$$

where $\mathbf{Q}_c$, $\mathbf{K}_c$, $\mathbf{V}_c$ are the query, key, and value matrices derived from $\mathbf{X}_c$, respectively, after linear projections. As shown in Fig. 4(d), $\hat{\mathbf{Y}}_c$ mainly captures narrowband information, i.e., low-frequency components that encode holistic writing characteristics.

Medium/Fine (Broadband/Extended-band) Branches. The medium and fine branches focus on fine-grained stroke details, such as stroke continuity (*e.g.*, short or elongated) and orientation transitions (*e.g.*, angular or smooth). To enhance local feature perception, the medium branch applies a 1×1 convolution followed by a depthwise convolution [42], and the fine branch applies max pooling (with a stride of 1 to preserve spatial resolution) followed by a 1×1 convolution. Both branches use GELU activation [19], formulated as:

$$\begin{aligned}\hat{\mathbf{Y}}_m &= \text{GELU}\big(\text{DwConv}(\text{PwConv}(\mathbf{X}_m))\big),\\ \hat{\mathbf{Y}}_f &= \text{GELU}\big(\text{PwConv}(\text{MaxPool}(\mathbf{X}_f))\big).\end{aligned} \tag{2}$$

Fig. 4(d) shows that $\hat{\mathbf{Y}}_m$ and $\hat{\mathbf{Y}}_f$ capture broadband and extended-band information, emphasizing relatively higher-frequency information compared with $\hat{\mathbf{Y}}_c$.

Intuitively, the medium branch captures mid-frequency variations, while the fine branch focuses on high-frequency details, encoding fine-grained stylistic cues.

Integration. The outputs from three branches are integrated via channel-wise concatenation:

$$\hat{\mathbf{Y}} = \text{Concat}(\hat{\mathbf{Y}}_c, \hat{\mathbf{Y}}_m, \hat{\mathbf{Y}}_f). \tag{3}$$

In addition, as shown in Fig. 4(a), we apply *Layer Normalization* (LN) and a *Multi-Layer Perceptron* (MLP) with a residual connection to enhance the fused feature. Therefore, the feature of a single sample is represented as:

$$\mathbf{Y} = \text{MLP}(\text{LN}(\mathbf{X} + \hat{\mathbf{Y}})) + \mathbf{X} + \hat{\mathbf{Y}}. \tag{4}$$

Through these progressive multi-band branches, FASE naturally emphasizes hierarchical stylistic cues from low to high frequencies, enabling fine-grained style capture for effective generation.

Contrastive Learning. To enforce style alignment and inter-style discrimination, we introduce a *triplet contrastive loss* [25]:

$$\mathcal{L}_{ctl} = -\log \frac{\exp(\mathbf{Y}_a \cdot \mathbf{Y}_p/\tau)}{\exp(\mathbf{Y}_a \cdot \mathbf{Y}_p/\tau) + \sum_{i=1}^{k} \exp(\mathbf{Y}_a \cdot \mathbf{Y}_{n_i}/\tau)}, \tag{5}$$

where $\mathbf{Y}_a$, $\mathbf{Y}_p$, and $\mathbf{Y}_{n_i}$ are all extracted from FASE, denoting the anchor (generated), positive (same style), and negative (different styles) samples, respectively; k is the number of negatives; $\tau = 0.07$ is the temperature hyperparameter; and all samples share the same character to isolate style from content.

3.4 Hierarchical Style Aggregation (HiSA)

To achieve fine-grained and coherent style imitation, we introduce the *Hierarchical Style Aggregation* (HiSA) mechanism, which explicitly transfers and fuses style cues in a coarse-to-fine manner, guided by the hierarchical composition of Chinese characters.

Radical/Component-level Style Transfer. At the component and radical levels, which encapsulate coarser structural information with greater stylistic variation, we employ an attention mechanism for style fusion. The continuous structural representations of the target character, denoted as $\mathbf{E}_c$ (component-level) and $\mathbf{E}_r$ (radical-level), are derived from their respective structural decomposition data $\mathcal{C}^{tgt}$ and $\mathcal{R}^{tgt}$. These representations are fused with the reference style features $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_k]$ using the following attention operation:

$$\mathbf{Z}_{r/c} = \text{softmax}\left(\frac{\mathbf{E}_{r/c}\mathbf{Y}^\top}{\sqrt{d_k}}\right)\mathbf{Y} \tag{6}$$

This process of cross-attention between target structure and reference style enables multi-granularity style aggregation, yielding two complementary stylized representations: $\mathbf{Z}_r$ at the radical level and $\mathbf{Z}_c$ at the component level.

Stroke-level Style Transfer. Since stroke-level features reflect fine-grained stylistic details, we apply similarity-based aggregation to enable precise alignment and selective feature fusion from structurally similar strokes. As illustrated in Fig. 3, we first map $\mathcal{S}^{tgt}$ and each $\mathcal{S}_i^{ref}$ into continuous representations $\mathbf{E}^{tgt} = \{\mathbf{e}_n\}_{n=1}^{n_s}$ and $\mathbf{E}_i^{ref} = \{\mathbf{e}_{i,j}\}_{j=1}^{m_i}$, respectively, where $i = 1, \ldots, k$.

Cosine similarity between each $\mathbf{E}_i^{ref}$ and $\mathbf{E}^{tgt}$ is used to quantify the correspondence between the target and reference stroke-level features:

$$\alpha_i = \frac{1}{m_i n_s} \sum_{j=1}^{m_i} \sum_{n=1}^{n_s} \frac{\mathbf{e}_{i,j}^\top \mathbf{e}_n}{\|\mathbf{e}_{i,j}\| \|\mathbf{e}_n\|}, \tag{7}$$

where n_s and m_i are the total number of strokes in the target character and the i-th reference glyph, respectively.

The stroke-aware aggregated style representation $\mathbf{Z}_s$ is computed via a normalized weighted sum:

$$\mathbf{Z}_s = \sum_{i=1}^{k} \xi_i \mathbf{Y}_i, \quad \text{with} \quad \xi_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{k} \exp(\alpha_j)}. \tag{8}$$

Aggregation. Finally, the overall multi-level style representation is formed by channel-wise concatenation:

$$\mathbf{Z} = \text{Concat}(\mathbf{Z}_s, \mathbf{Z}_r, \mathbf{Z}_c). \tag{9}$$

$\mathbf{Z}$ aggregates structural and style features as the condition input to an online or offline handwriting decoder for generating the target character.

3.5 Versatile Handwriting Generation

Online Handwriting Generation. We employ a Transformer decoder [50] attending to $\mathbf{Z}$ to generate handwriting trajectories. The component-level representation $\mathbf{E}_c$ is used as the initial token of the decoder, providing structural cues for trajectory generation. At each time step t, the output follows a *Gaussian Mixture Model* (GMM) [17] consisting of R bivariate normal distributions, which captures the uncertainty and multimodality of pen-tip movements:

$$\mathbf{o}_t = (\{\hat{\pi}^r, \hat{\mu}_x^r, \hat{\mu}_y^r, \hat{\sigma}_x^r, \hat{\sigma}_y^r, \hat{\rho}_{xy}^r\}_{r=1}^{R}, \hat{q}_1, \hat{q}_1, \hat{q}_1), \tag{10}$$

where $\hat{\pi}^r$ denotes the weight of the r-th Gaussian component, $\hat{\mu}_x^r$ and $\hat{\mu}_y^r$ represent the means of the distributions, $\hat{\sigma}_x^r$ and $\hat{\sigma}_y^r$ denote their standard deviations, and $\hat{\rho}_{xy}^r$ represents the correlation coefficient. Finally, $\hat{q}_1$, $\hat{q}_2$, and $\hat{q}_3$ correspond to the predicted pen states (down, up, end), respectively.

For trajectory optimization, we minimize both the negative log-likelihood loss of the trajectory points and the cross-entropy loss for pen state prediction:

$$\mathcal{L}_{\text{GMM}} = -\frac{1}{T} \sum_{t=1}^{T} [\log p(\Delta x_t, \Delta y_t) + \lambda_{\text{state}} \sum_{i=1}^{3} q_{t,i} \log \hat{q}_{t,i}], \tag{11}$$

where T denotes the number of trajectory points, and $\lambda_{\text{state}} = 2$ is the hyperparameter. The probability density $p(\Delta x, \Delta y)$ is modeled by a GMM: $p(\Delta x, \Delta y) =$

$\sum_{r=1}^{R} \hat{\pi}^r \mathcal{N}(\Delta x, \Delta y | \hat{\mu}_x^r, \hat{\mu}_y^r, \hat{\sigma}_x^r, \hat{\sigma}_y^r, \hat{\rho}_{xy}^r)$, where $\mathcal{N}(\cdot)$ is the bivariate normal distribution function.

For *online* (trajectory-based) handwriting generation, the total loss is:

$$\mathcal{L}_{\text{online}} = \lambda_{\text{ctl}} \mathcal{L}_{\text{ctl}} + \lambda_{\text{GMM}} \mathcal{L}_{\text{GMM}}, \tag{12}$$

where λ_{ctl} and λ_{GMM} are weighting parameters and set to 1 by default.

Offline Handwriting Generation. We adopt a DDPM-based conditional model [21] for offline handwriting generation, where the style representation $\mathbf{Z}$ is injected into the U-Net via cross-attention to guide the denoising process. The training objective is to explicitly minimize the mean squared error between the predicted noise $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{Z})$ and the ground-truth noise ϵ:

$$\mathcal{L}_{\text{MSE}} = \left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{Z}) \right\|^2. \tag{13}$$

Besides, we incorporate a perceptual loss [23]:

$$\mathcal{L}_{\text{VGG}} = \sum_{l} ||\text{VGG}^{(l)}(\mathbf{x}_0) - \text{VGG}^{(l)}(\mathbf{x}_{\text{target}})||, \tag{14}$$

where $\text{VGG}^{(l)}(\cdot)$ denotes the features extracted from the l-th layer of a pretrained VGG16 network [43], and we use layers *relu*1_2, *relu*2_2 and *relu*3_3.

For *offline* (image-based) handwriting generation, the total loss becomes:

$$\mathcal{L}_{\text{offline}} = \lambda_{\text{ctl}} \mathcal{L}_{\text{ctl}} + \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}}, \tag{15}$$

where λ_{ctl}, λ_{MSE} and λ_{VGG} are hyperparameters, with $\lambda_{\text{VGG}} = 0.01$, and all others set to 1 by default.

4 Experiments

4.1 Settings

Handwriting Datasets. We train our model on the CASIA-OLHWDB (1.0–1.2) and CASIA-HWDB (1.0–1.2) [29] datasets for online and offline handwriting generation, respectively. Both datasets contain approximately 3.7 million handwritten Chinese character samples collected from 1,020 writers, provided as trajectory sequences and images. For evaluation, we use both the online and offline versions of the ICDAR 2013 competition dataset [56], which contain handwriting data from 60 writers, each writing 3,755 commonly used Chinese characters. For online handwriting generation, trajectories are rendered into images. All images are cropped and resized to 64×64.

Evaluation Metrics. For online character generation, we employ *Dynamic Time Warping* (DTW) [3, 6] to measure the temporal alignment between predicted and target trajectories. In addition, we report *Style Score* [48] and *Content Score* [59], computed using pretrained style classification and character recognition models to evaluate stylistic fidelity and content accuracy, respectively. For

Table 1: Quantitative comparison on *online* Chinese handwriting generation.

Method	Venue	DTW↓	Style↑	Content↑
DeepImitator [59]	PR'20	1.0476	62.56	92.07
WriteLikeYou [48]	CGF'21	0.9244	84.78	94.26
SDT [9]	CVPR'23	<u>0.8780</u>	<u>94.56</u>	<u>96.47</u>
OLHWG [40]	ICLR'25	0.9173	91.28	90.11
DNA [22]	WACV'26	0.9041	93.43	96.23
HiSAC (Ours)	-	**0.8224**	**97.01**	**96.68**

offline character generation, we adopt several widely used image quality metrics, including *Fréchet Inception Distance* (FID) [20], *Learned Perceptual Image Patch Similarity* (LPIPS) [57], and *Root Mean Square Error* (RMSE) between generated and ground-truth images. We also use pretrained recognizers to further assess the consistency of style and content.

4.2 Comparison with Online Methods

We compare HiSAC with several state-of-the-art methods for online handwriting generation, including DeepImitator [59], WriteLikeYou [48], OLHWG [40], SDT [9], and DNA [22]. Notably, OLHWG is designed for text-line-level generation. To ensure a fair comparison, we replace the original RNN or LSTM backbones of DeepImitator and WriteLikeYou with Transformer architectures and retrain them under the same experimental settings. For WriteLikeYou, we follow SDT and adopt a CNN-Transformer architecture for both content and style encoding.
Quantitative Comparison. Table 1 presents quantitative results on the ICDAR 2013 competition dataset. Our method outperforms all compared approaches in trajectory accuracy, style similarity and content fidelity. Specifically, compared with the second-best method, HiSAC reduces DTW by 0.0556 and improves Style Score by 2.45, demonstrating its ability to generate online handwriting that is both precise and exhibits excellent style imitation performance.
Qualitative Comparison. As shown in Fig. 5, DeepImitator struggles to imitate the writing style and suffers from structural collapse in some generated samples. In comparison, WriteLikeYou, SDT and DNA better imitate the writing style but fall short in capturing fine details. Conversely, HiSAC excels in style imitation, achieving both faithful style replication and structural integrity.

4.3 Comparison with Offline Methods

We compare the offline variant of our model with existing offline handwriting generation methods, including the GAN-based HWT [4] and VATr [39], as well as the diffusion-based WordStylist [33], DiffusionPen [34], One-DM [7], and DiffBrush [8]. As some of these methods are originally tailored for English or Latin scripts and are less capable of modeling the structural complexity of Chinese characters, we incorporate the SDT content encoder, which is specifically designed for Chinese handwriting. Following SDT and DiffusionPen, for methods originally designed for one-shot generation, such as One-DM, we extract style

Fig. 5: Qualitative comparisons on both *online* and *offline* Chinese handwriting generation tasks with SOTA methods.

features from five reference images and average them to ensure a consistent reference setting across all baselines. In addition, we implement both diffusion-based and VQ-VAE+Transformer autoregressive (VAR) [44] generators to demonstrate that HiSAC is *generator-agnostic* (Table 2). Unless otherwise specified, all subsequent analyses are conducted using the diffusion-based generator.

Quantitative Comparison. As shown in Table 2, our method also demonstrates a significant advantage in offline generation. GAN-based models such as HWT and VATr are originally designed for English scripts, making them less effective when transferred to Chinese character generation; notably, HWT's reliance on string embeddings as content input limits its Chinese generation performance. Although diffusion-based methods yield similar scores in image quality metrics, our model consistently achieves superior performance in both content fidelity and style consistency.

Qualitative Comparison. As shown in Fig. 5, HWT and VATr struggle to handle the structural complexity of Chinese characters. WordStylist, due to its reliance on WriterID embeddings, lacks the capability for flexible style imitation.

Table 2: Quantitative comparison on *offline* Chinese handwriting generation.

Method	Venue	FID↓	LPIPS↓	RMSE↓	Style↑	Content↑
HWT [4]	ICCV'21	162.43	0.3112	0.3245	3.78	5.23
VATr [39]	CVPR'23	136.78	0.2733	0.3105	38.49	43.25
WordStylist [33]	ICDAR'23	39.27	0.1725	0.1661	57.74	82.79
DiffusionPen [34]	ECCV'24	35.14	0.1678	0.1608	61.78	86.18
One-DM [7]	ECCV'24	35.45	0.1652	0.1520	63.47	86.27
DiffBrush [8]	ICCV'25	36.41	0.1704	0.1573	66.18	87.11
$\text{HiSAC}_{\text{Diffusion}}$ (Ours)	-	34.88	**0.1627**	0.1531	**68.21**	**89.48**
$\text{HiSAC}_{\text{VAR}}$ (Ours)	-	**34.17**	0.1638	**0.1485**	68.04	89.22

Table 3: Ablation study on FASE and HiSA. ✓: used; ✗: removed.

FASE			**HiSA**					
Coarse	*Medium*	*Fine*	*Stroke*	*Component*	*Radical*	DTW↓	Style↑	Content↑
✓	✗	✗	✓	✓	✓	0.8318	93.37	96.49
✗	✓	✗	✓	✓	✓	0.8286	93.75	96.58
✗	✗	✓	✓	✓	✓	0.8254	94.50	**96.75**
✓	✓	✓	✗	✗	✗	1.2146	83.22	67.35
✓	✓	✓	✗	✓	✓	0.8373	96.88	95.62
✓	✓	✓	✓	✗	✓	0.8321	96.62	96.18
✓	✓	✓	✓	✓	✗	0.8410	96.91	95.80
✓	✓	✓	✓	✓	✓	**0.8224**	**97.01**	96.68

DiffusionPen, as an improved version of WordStylist, supports arbitrary style generation; however, it performs poorly in handling fine-grained details. One-DM and DiffBrush further enhance the style encoder, achieving strong style imitation performance, and our model reaches a comparable qualitative performance.

4.4 Ablation and Analysis

We perform ablation experiments to validate the core architectural designs of HiSAC. For brevity, these studies are primarily conducted in the online setting to emphasize structural and trajectory accuracy. Additional results for offline tasks are included in the Supplementary Material due to space constraints.

Effectiveness of Different Modules. As shown in Table 3, we conduct ablation studies to assess the contribution of each module. Removing any one of the coarse, medium, or fine branches decreases Style Score, while combining all leads to clear improvement, showing the importance of capturing global structure and fine-grained details. Excluding HiSA significantly reduces Content Score to 67.35, as the lack of structural guidance leads to poor character composition. Moreover, removing any level—stroke, component, or radical—consistently harms performance, confirming that multi-level structural cues are essential for accurate handwriting generation.

Multi-band Style Feature Visualization. We further visualize the output features and corresponding frequency spectra from multi-band branches in FASE. As shown in Fig. 6, the coarse (narrowband) branch predominantly encodes low-frequency components, capturing holistic writing characteristics. The medium (broadband) branch focuses more on mid-frequency variations, while the fine

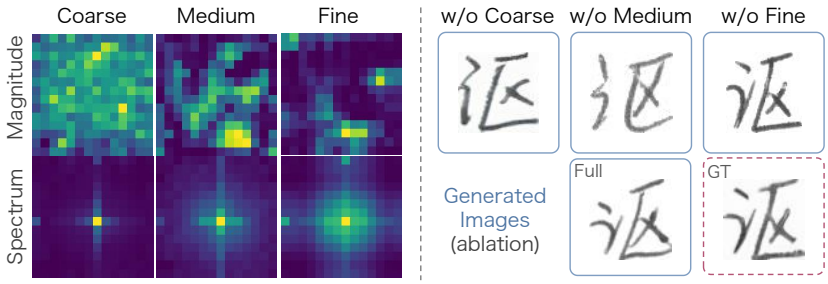

Fig. 6: Ablation study on the frequency-aware multi-band style encoding (FASE). *Left*: channel-wise average magnitudes and their frequency spectrum of multi-band feature maps. *Right*: generated results.

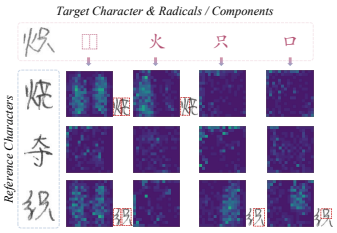

Fig. 7: Attention visualization.

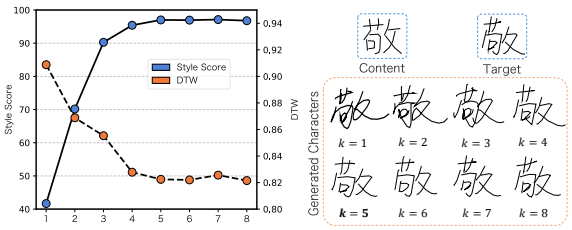

Fig. 8: Effect of the number of reference glyphs.

(extended-band) branch emphasizes high-frequency details, encoding fine-grained stylistic cues. Note that FASE does not enforce a strict, non-overlapping frequency decomposition; thus, some spectral overlap is expected, and the fine branch may also respond to low-to-mid frequency regions, partly because surrounding blank regions introduce low-frequency responses. Moreover, by removing individual branches, we visualize the resulting handwriting changes, which clearly demonstrate that each branch contributes divergent and complementary information to the final generation.

Attention (Correspondence) Visualization. We visualize the attention maps at the component and radical levels. Given a component or radical in the target character as the query, we extract its corresponding attention map from the spatial correspondence matrix (Eq. 6). As shown in Fig. 7, when reference images contain matching components or radicals, the attention maps focus more on these shared elements, while images without such components receive less attention. This demonstrates that HiSA can learn precise cross-scale matching relationships between target characters and reference glyphs, thereby improving the quality of generated handwriting.

Effect of the Number of Reference Glyphs. We investigate the effect of different numbers of reference samples, which significantly influences the model's performance. As shown in Fig. 8, when k exceeds 5, both Style Score and DTW tend to stabilize. In addition, the visualized handwriting results under varying k values further reflect this trend—as k increases, the generated characters become progressively more stable in structure and style. Based on this observation, we

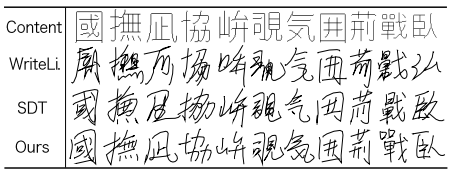

Fig. 9: Zero-shot generation of Japanese Characters (without training).

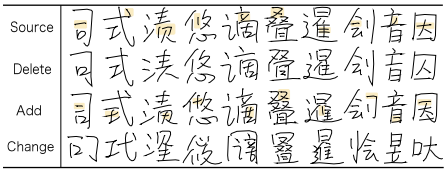

Fig. 10: Application to the character editing task.

set k to 5 in all experiments. It is worth mentioning that all compared methods also adopt five reference samples [7, 8], ensuring a completely fair comparison.

Generation with References Sharing Target Radicals. While the main experiments use randomly selected references to ensure unbiased evaluation, we further deliberately choose reference characters sharing common radicals with the targets to better verify the capability of FASE and HiSA for precise style transfer. As shown in Fig. 13, our model accurately localizes corresponding radical regions and transfers fine-grained style features.

4.5 Generalization Experiments

Beyond standard handwriting benchmarks, we further evaluate the versatility and robustness of HiSAC across several challenging extensions.

Zero-shot Generation of Japanese Characters. Japanese Kanji share substantial structural similarities with Chinese characters in terms of compositional logic and basic components, making them suitable for evaluating cross-domain generalization. We therefore compare HiSAC against SDT [9] and WriteLikeYou [48] on generating unseen characters, covering both traditional Chinese characters and Japanese Kanji. As shown in Figure 9, these dual-encoder methods often suffer from missing or redundant strokes and structural collapse, suggesting that their content encoders fail to capture how characters should be written. In contrast, our model generalizes more effectively to novel structures.

Application to Character Editing. We modify the three-level hierarchical input by editing strokes and adjusting spatial layout. As shown in Fig. 10, these changes are faithfully reflected in the generated glyphs, demonstrating that HiSAC successfully learns fine-grained structural information and synthesizes characters consistent with the modified input.

Application to Artistic Fonts. We apply our model to Chinese artistic font generation, a task that differs from natural handwriting as artistic fonts often exhibit exaggerated deformations and decorative variations. To handle these stylistic characteristics, we retrain the model on an artistic font dataset and compare its performance with CF-Font (CF) [51] and FontDiffuser (FD) [55]. As shown in Fig. 11 and Table 4, our model achieves competitive results, demonstrating its effectiveness in generating high-quality and stylistically faithful artistic fonts.

Application to English Handwriting. We adapt our model to online English handwriting generation on the BRUSH dataset [27] by treating each letter as a radical arranged in a fixed left-to-right layout. Although English lacks the

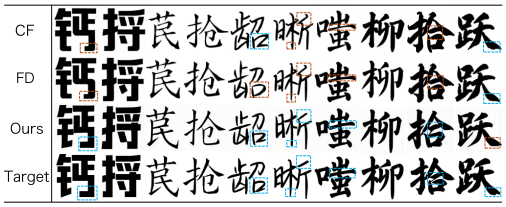

Fig. 11: Comparisons on Chinese artistic font generation with CF-Font and FontDiffuser.

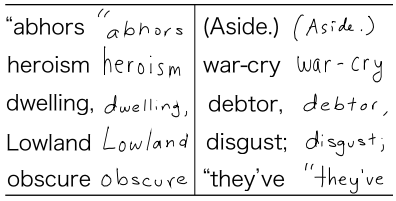

Fig. 12: Generalization to English handwriting generation.

Table 4: Quantitative comparisons on Chinese artistic font generation.

Method	Venue	FID↓	LPIPS↓	SSIM↑	L1↓
CF-Font [51]	CVPR'23	19.27	0.1601	0.4259	0.1796
FontDiffuser [55]	AAAI'24	**14.84**	0.1486	0.4418	0.1715
HiSAC	–	16.23	**0.1395**	**0.4684**	**0.1655**

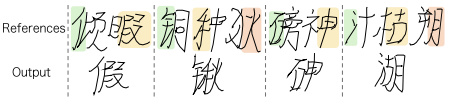

Fig. 13: Generation results with references sharing target radicals.

hierarchical compositional structure of Chinese, HiSAC still produces visually coherent and legible word sequences (Fig. 12). This cross-lingual adaptation further demonstrates the model's flexibility and its ability to generalize to alphabetic scripts with entirely different writing systems.

5 Conclusion

In this paper, we propose HiSAC, a composition-guided hierarchical style aggregation framework for versatile Chinese handwriting generation. By leveraging multi-granularity information in both spatial (compositional structure) and frequency (style characteristics) domains, HiSAC significantly enhances structural and stylistic consistency, and demonstrates strong generalization capability to out-of-vocabulary and cross-language scripts. However, learning efficient representations for characters with extremely complex structures or severe stylistic deformations (*e.g.*, running script or cursive handwriting) remains a challenging direction for future work. Besides, the proposed hierarchical representation learning and compositional structure modeling hold significant promise for document analysis and understanding. We will explore these issues in future work.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096 and 62571395, in part by Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Establishment Fund of the State Key Laboratory of Wakefulness/Sleep and Cognition (Chongqing Institute for Brain and Intelligence), and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138.

References

1. Aksan, E., Pece, F., Hilliges, O.: Deepwriting: Making digital ink editable via deep generative modeling. In: ACM CHI. 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. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: ACM SIGKDD Workshop. pp. 359–370 (1994)
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. Chen, L., Gu, L., Fu, Y.: Frequency-dynamic attention modulation for dense prediction. In: ICCV. pp. 22620–22632 (2025)
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., Ke, Q., Guo, Q., Huang, S.: One-shot diffusion mimicker for handwritten text generation. In: ECCV. pp. 410–427 (2024)
8. Dai, G., Zhang, Y., Qin, Y., Guo, Q., Huang, S., Yan, S.: Beyond isolated words: Diffusion brush for handwritten text-line generation. In: ICCV. pp. 19054–19064 (2025)
9. 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)
10. 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)
11. 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)
12. Dong, Y., Cordonnier, J.B., Loukas, A.: Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: ICML. pp. 2793–2803. PMLR (2021)
13. 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)
14. Fu, B., He, J., Wang, J., Qiao, Y.: Neural transformation fields for arbitrary-styled font generation. In: CVPR. pp. 22438–22447 (2023)
15. Gan, J., Wang, W.: Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles. In: AAAI. pp. 7484–7492 (2021)
16. Gan, J., Wang, W., Leng, J., Gao, X.: Higan+: Handwriting imitation gan with disentangled representations. ACM TOG **42**(1), 1–17 (2022)
17. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
18. Gui, D., Chen, K., Ding, H., Huo, Q.: Zero-shot generation of training data with denoising diffusion probabilistic model for handwritten chinese character recognition. In: ICDAR. pp. 348–365. Springer (2023)
19. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)

21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS. vol. 33, pp. 6840–6851 (2020)
22. Huang, T.L., Do-Tran, N.T., Le, N.H.L., Shuai, H.H., Huang, C.C.: Dna: Dual-branch network with adaptation for open-set online handwriting generation. In: WACV. pp. 4170–4179 (2026)
23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711. Springer (2016)
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. Kong, Y., Luo, C., Ma, W., Zhu, Q., Zhu, S., Yuan, N., Jin, L.: Look closer to supervise better: one-shot font generation via component-based discriminator. In: CVPR. pp. 13482–13491 (2022)
27. Kotani, A., Tellex, S., Tompkin, J.: Generating handwriting via decoupled style descriptors. In: ECCV. pp. 764–780. Springer (2020)
28. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: CVPR. pp. 26763–26773 (2024)
29. Liu, C.L., Yin, F., Wang, D.H., Wang, Q.F.: Casia online and offline chinese handwriting databases. In: ICDAR. pp. 37–41 (2011)
30. Liu, Y., Khalid, F.B., Wang, L., Zhang, Y., Wang, C.: Elegantly written: Disentangling writer and character styles for enhancing online chinese handwriting. In: ECCV. pp. 409–425 (2025)
31. Mattick, A., Mayr, M., Seuret, M., Maier, A., Christlein, V.: Smartpatch: Improving handwritten word imitation with patch discriminators. In: ICDAR. pp. 268–283. Springer (2021)
32. Nguyen, T., Nguyen, T., Baraniuk, R.: Mitigating over-smoothing in transformers via regularized nonlocal functionals. NeurIPS pp. 80233–80256 (2023)
33. 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)
34. Nikolaidou, K., Retsinas, G., Sfikas, G., Liwicki, M.: Diffusionpen: Towards controlling the style of handwritten text generation. arXiv preprint arXiv:2409.06065 (2024)
35. Pan, W., Zhu, A., Zhou, X., Iwana, B.K., Li, S.: Few shot font generation via transferring similarity guided global style and quantization local style. In: ICCV. pp. 19506–19516 (2023)
36. Park, N., Kim, S.: How do vision transformers work? In: ICLR (2022)
37. Park, S., Chun, S., Cha, J., Lee, B., Shim, H.: Few-shot font generation with localized style representations and factorization. In: AAAI. vol. 35, pp. 2393–2402 (2021)
38. Park, S., Chun, S., Cha, J., Lee, B., Shim, H.: Multiple heads are better than one: Few-shot font generation with multiple localized experts. In: ICCV. pp. 13900–13909 (2021)
39. Pippi, V., Cascianelli, S., Cucchiara, R.: Handwritten text generation from visual archetypes. In: CVPR. pp. 22458–22467 (2023)
40. Ren, M., Zhang, Y.M., Chen, Y.: Decoupling layout from glyph in online chinese handwriting generation. In: ICLR. pp. 4716–4733 (2025)

41. Ren, Z., Pan, Y., Chen, J., Zhao, L., Liao, M., Qian, X., Gong, W.: A survey on deep learning-based chinese font style transfer. IEEE TAI (2025)
42. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
44. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
45. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
46. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
47. Tang, L., Cai, Y., Liu, J., Hong, Z., Gong, M., Fan, M., Han, J., Liu, J., Ding, E., Wang, J.: Few-shot font generation by learning fine-grained local styles. In: CVPR. pp. 7895–7904 (2022)
48. Tang, S., Lian, Z.: Write like you: Synthesizing your cursive online chinese handwriting via metric-based meta learning. Computer Graphics Forum **40**(2), 141–151 (2021)
49. Tang, S., Xia, Z., Lian, Z., Tang, Y., Xiao, J.: Fontrnn: Generating large-scale chinese fonts via recurrent neural network. In: CGF. pp. 567–577 (2019)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
51. Wang, C., Zhou, M., Ge, T., Jiang, Y., Bao, H., Xu, W.: Cf-font: Content fusion for few-shot font generation. In: CVPR. pp. 1858–1867 (2023)
52. Wang, L., Wang, C., Liu, Y.: Ehw-font: A handwriting enhancement approach mimicking human writing processes. ESWA **278**, 127278 (2025)
53. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. arXiv preprint arXiv:2203.05962 (2022)
54. Xie, Y., Chen, X., Sun, L., Lu, Y.: DG-Font: Deformable generative networks for unsupervised font generation. In: CVPR. pp. 5130–5140 (2021)
55. Yang, Z., Peng, D., Kong, Y., Zhang, Y., Yao, C., Jin, L.: Fontdiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In: AAAI. pp. 6603–6611 (2024)
56. Yin, F., Wang, Q.F., Zhang, X.Y., Liu, C.L.: Icdar 2013 chinese handwriting recognition competition. In: ICDAR. pp. 1464–1470 (2013)
57. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: CVPR. pp. 586–595 (2018)
58. Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y.: Drawing and recognizing chinese characters with recurrent neural network. IEEE TPAMI **40**(4), 849–862 (2017)
59. 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)
60. Zhu, Y., Li, Z., Wang, T., He, M., Yao, C.: Conditional text image generation with diffusion models. In: CVPR. pp. 14235–14245 (2023)