Write Freely: Disentangling Content and Style for Multi-Scale Autoregressive Reconstruction of Online Handwriting Trajectories

Yu Liu[1,2,†], Huawei Qiu[1], Chao Liu[3,†], Yang Ding[4], Yuqiu Kong[1], Cunrui Wang[1,2,*]

[1] Dalian Chinese Font Design Technology Innovation Center, Dalian Minzu University, Dalian, China
[2] Key Laboratory of Education Informatization for Nationalities (Yunnan Normal University), Ministry of Education,Yunnan, China
[3]Inner Mongolia Normal University, Hohhot, China
[4]Changchun University of Science and Technology, Changchun, China

ethanliuyu@foxmail.com, 1227009932@qq.com, 2012392967@qq.com,

2023200170@mails.cust.edu.cn, {yqkong,crw}@dlnu.edu.cn

Abstract

Although electronic writing tools and environments enhance writing efficiency and convenience, the absence of the mechanical feedback between pen tip and medium can easily lead to jitter, unintended joins, and structural deviations, thereby weakening character legibility and personalized style. To address this, this paper proposes a handwriting trajectory reconstruction method that decouples and recombines content and style. To systematically learn the dynamic process of human writing and achieve deep integration with device-side rendering while retaining editability, we represent handwritten strokes as sequences of discrete coordinates. Inspired by the human hierarchical writing and perception process of "global first, local later," we model handwriting trajectory reconstruction as a multiscale progressive refinement task. We divide the reconstruction process into two stages. In the content preservation stage, original handwriting drawing parameters are used as content guidance to maintain the consistency of character content. In the style aggregation stage, inspired by representing style via amplitude components, we propose a phase-frozen amplitude-perturbation separation scheme to achieve style control. Experiments show that the proposed method not only improves character legibility and structural stability but also better preserves individualized handwriting style, generating more natural stroke details compared to single-scale autoregressive and image re-rendering approaches.

Keywords: Handwriting Imitation, Handwriting Character Optimization, Path Reconstruction, Autoregressive Models, Next-Scale Prediction, Frequency Learning

1. Introduction

Since humans mastered writing, handwriting has always been the foundation of information recording and transmission. In the digital era, although electronic writing tools have greatly improved efficiency and convenience, the lack of tactile feedback and pen resistance on screens causes handwriting to wobble, disconnect, or become incomplete during rapid writing, diminishing readability and individual writing style. Conversely, if one aims for neatness and stylistic expression, speed must be reduced, sacrificing efficiency. Therefore, optimizing electronic handwriting trajectories can enhance character readability while effectively preserving the personalized features of users' handwriting during high-speed writing.

The current optimization for electronic handwritten characters mainly involves recognizing the handwriting first [2, 9, 43, 8] and then replacing it with character glyphs from a font library for display. This process can improve legibility, but users prefer their personalized handwriting to be naturally preserved. A strategy that balances both is to treat the conversion from original handwriting to optimized handwriting as an image-to-image mapping of handwritten Chinese character fonts [21, 22, 23]. The model learns style embeddings from the user's handwritten images and uses them as conditional inputs to a generator, which re-renders and reproduces the characters, resulting in handwritten character images that are both more readable and retain personal style. Although image generation models can reconstruct handwritten character images, static image representations are difficult to capture dynamic features such as stroke order and speed, lack cross-resolution reuse capabilities, and are challenging to integrate deeply with device-

*† Equal Contribution ⋆ Corresponding Author

side rendering [37, 11].

In order to learn the dynamic process of human writing more systematically and achieve deep integration with device-side rendering while retaining editability, some studies propose a method based on sequence model to represent handwriting character handwriting as a writing track coordinate sequence, and use RNN, LSTM or Transformer pair sequences to predict the writing track coordinate sequence in an autoregressive way. The autoregressive model captures the dynamic process of human writing to a certain extent by predicting the next coordinate [24] in the sequence in a given context, but in the handwriting trajectory generation task, the single-scale autoregressive method has the accumulation of errors on the one hand, and on the other hand, it cannot take into account the overall structure and stroke details of Chinese characters in each step of prediction, resulting in local smoothing of the generated results and the imbalance of the overall structure [25, 38]. In order to alleviate this contradiction, one approach is to use the diffusion model to approximate the trajectory distribution layer by layer with multi-step denoising iteration. Another method is to disassemble the reconstruction process into multiple stages such as skeleton type, stroke refinement and dynamic parameter optimization, so that the global structure and local details can be optimized separately [34, 14]. Although the aforementioned method alleviates the problem to some extent, the real writing process is not a simple stack of immediate reactions. Before the writer writes, he usually has formed the overall outline of the target word in his memory. Then, under the constraints of this framework, the specific trajectory is gradually refined and generated according to the stroke timing. Therefore, a top-down, multi-scale, coarse-to-fine cognitive mechanism is closer to the actual state of writing, and provides new enlightenment for handwriting trajectory modeling.

Handwriting naturally embodies randomness and heterogeneity driven jointly by the writer's habits and kinematic noise, making the reconstruction of handwritten characters a significantly ill-posed problem [20]. Between the offline image domain and the online trajectory (temporal coordinate) representation, style (such as stroke slant, ligature, and turns) and content (character structure) are highly coupled, causing style injection in style transfer to easily introduce systematic perturbations to content features [10, 25]. Studies have indicated that when analyzing image information in the frequency domain, the magnitude component primarily reflects rich textures and surface details, while the phase component consistently plays a key role in preserving image structure and semantic information across different image domains [6, 3, 40]. Based on this observation, we analyzed the magnitude and phase of handwriting trajectory coordinate sequences in the frequency domain. As shown in the Fig. 1, for the same writer, the magnitude spectra of different characters are similar, suggesting that the magnitude primarily captures the stylistic motion statistics of writing. Accordingly, we consider using the magnitude as the main carrier for style injection, while employing the phase as the core constraint to preserve content.

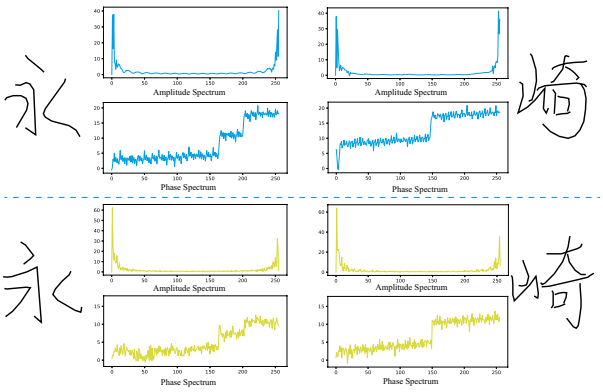

Figure 1. Example of frequency-domain analysis of handwritten trajectory coordinate sequences. Within the same writer, the amplitude spectra of different characters are similar, reflecting style consistency; in cross-writer handwriting, the phase spectra of the same characters are more consistent in abrupt change positions and overall trends.

For online handwriting scenarios, we conduct adaptive optimization of trajectories affected by stroke jitter, displacement, and random stroke connections, enhancing character legibility while faithfully preserving the writer's style. Given that the coordinate sequences of online handwriting trajectories more accurately reveal writing motion patterns and inherently offer resolution independence and easy end-side rendering integration, we decouple character content and writing style from the trajectories. Under any content-style combination, we reconstruct handwriting trajectory sequences using an autoregressive approach. We propose a multi-scale autoregressive reconstruction framework for discrete stroke sequences, first reconstructing the overall trajectory structure at the lowest scale, and then progressively completing stroke details at higher scales, achieving a coarse-to-fine trajectory reconstruction. We divide the reconstruction process into two stages: during the content-preserving stage, original handwriting drawing parameters serve as content guidance, maintaining handwriting content information in multi-scale autoregressive prediction. In the style-aggregation stage, amplitude carries style features while phase reveals temporal order and structural information; by fixing the phase components and learning the user's writing style while perturbing the amplitude components, trajectory style control is achieved. The contributions of this paper are summarized as follows:

- Propose a method for optimizing handwritten scripts based on a multi-scale autoregressive model. The model optimizes handwriting by learning writing

styles and processes from a small number of character samples provided by the user. The code is public at: https://github.com/ethanliuyu/WriteFreely

- Inspired by the human mechanism of writing and perception, which follows a "global before local" approach, we model the optimization of handwritten trajectories as a multi-scale, coarse-to-fine progressive refinement process.
- Propose content guidance based on "imperfect" handwriting, and through the display of injected geometric and temporal inductive biases, ensure that the generation process from coarse to fine continuously aligns the character skeleton with the writing sequence.
- Inspired by the representation of style through amplitude components, a scheme for separating frozen phase and perturbed amplitude is proposed to achieve style control.

2. Related Work

2.1. Font Generation

In recent years, Chinese character font stylization has become one of the research hotspots [37]. This task is typically viewed as an image-to-image translation problem, aiming to achieve the automatic generation and transfer of font styles. Some methods use an end-to-end approach to directly generate raster font images containing 9,169 characters [46, 45, 35, 21]; other studies employ unsupervised learning to separately encode the style and content of Chinese characters as independent representations, generating raster font images with specific style and content combinations [36, 27, 51, 47]. Additionally, some studies use sequence generation models [33, 20], representing vector glyphs as a series of drawing instruction sequences, and encoding and decoding these sequences using RNN , LSTM , or Transformer models to achieve stylized generation of Chinese fonts. Although existing research primarily focuses on the generation of printed brush fonts, the technical frameworks they rely on are fundamentally aligned with the handwritten stroke optimization task, providing theoretical support for this paper's exploration of Chinese handwritten style optimization.

2.2. Style Integration Strategy

Image-to-image (I2I) translation aims to learn the conditional mapping between the source and target domains, so that the output presents the appearance and style of the target domain while maintaining the content structure of the source image [48, 15]. In this framework, font generation can be considered a typical I2I task, using the glyph structure of the source font as a content constraint and mapping it to the target style space, resulting in results that match the target domain in terms of style attributes [21, 35]. Among them, the few-shot font generation (FFG) method has attracted widespread attention because it decouples font image content and style, and generates new fonts by combining arbitrary content and style features [22, 23]. According to the fusion mode of style feature characterization, the FFG method can be divided into two categories: global style coding injection and local structured style injection. Global style encoding injection encodes the overall style of the reference font into a global vector or multiscale feature map, and modulates the content features in the generator by stitching, AdaIN/FiLM, or across attention, achieving a style that gives the target domain while maintaining the source character skeleton [39, 36, 26]. Local structured style injection refines the style representation to the level of components or strokes, and selectively injects it into the corresponding content area in the way of component alignment and cross-attention, so as to solve the problem of consistency and local diversity of combined text across words [25, 41, 28]. However, this type of method will couple with the content representation and easily cause disturbances to the content structure regardless of whether it is style injection in the time domain or the image domain. In the latest research, frequency domain analysis is increasingly being incorporated into deep learning models. Some studies have pointed out that when analyzing image information in the frequency domain, the amplitude component mainly reflects rich texture and surface details, while the phase component always plays a key role in image structure and semantic information in different image domains [6, 3, 40]. Accordingly, this paper proposes a style fusion strategy in the Fourier domain, which realizes trajectory style regulation by disturbing the amplitude component under the condition of fixed phase component, and ensures that the structure and semantic information of the trajectory remain unchanged.

2.3. Handwriting Font Generation

Compared with standardized typesetting fonts, handwriting is affected by human hand movements and habits, and occasional ups and downs and individual differences are inevitable in the writing process, making its reconstruction a difficult task [20, 18]. Based on capture technology, there are two main methods of obtaining handwriting information: offline and online. For offline methods, static 2D images captured by image scanning devices. The stylized reconstruction task of handwritten characters is represented as an image style transfer task, and the style of the target domain is integrated while maintaining the content of the source image by learning the mapping function between the source domain and the target domain, so as to realize the stylized modeling of the handwritten font [13, 29, 44].

However, unlike the static image paradigm of overall rendering, human writing is a step-by-step formative process constrained by stroke order, accompanied by the temporal evolution of speed and trajectory. Therefore, on the one hand, it is difficult to characterize the timing dynamics and stroke dependencies of pure static image representation, and on the other hand, it is difficult to integrate with on-device rendering due to the lack of cross-resolution multiplexing capabilities. For online methods, additional dynamic information of the handwriting can be captured through special input devices such as stylus, stylus, etc., including writing speed, pressure, tilt angle, etc., and the coordinate information of the nib trajectory can also be utilized. In order to characterize the timing characteristics of "write-out", a generative model based on vector representation with topological information is proposed [7, 1]. This method encodes the process signals such as the drawing direction and stroke order of the handwriting into a drawing instruction sequence, and learns its conditional probability distribution to reproduce dynamic writing. Given the variable length of vector sequences, most models adopt an autoregressive structure [31, 1] to naturally handle vector drawing parameters of variable length under a unified framework [25, 30, 16]. Based on the advantages of online handwriting being able to explicitly characterize dynamic features such as stroke order and speed, and vector instructions naturally support cross-resolution multiplexing and device-side rendering integration, this paper uniformly represents handwriting as a discrete coordinate sequence for modeling.

2.4. Autoregressive modeling

Autoregressive (AR) modeling [4] has been widely used in the generative field, typically discretizing successive signals into token sequences and then predicting the next token [12, 42] under the condition of a given prefix. This route has been a remarkable success in text. However, the image [34] and handwritten character trajectories [25] present two-dimensional spatial dependence and time dimension coupling at the same time, and the traditional AR assumption of "one-dimensional causal order" is difficult to take into account the multi-dimensional structure and temporal constraints in a single inference, which can easily lead to the local destruction of the structure, insufficient long-distance dependent characterization, and error accumulation [17]. In order to break through this bottleneck, the recently proposed Visual Autoregressive Modeling (VAR) redefines the AR learning of images from "next token prediction" to "next scale (resolution) prediction", and generates multi-scale discrete token graphs from top to bottom, from coarse to fine. Tokens within the same scale can be generated in parallel, so as to preserve spatial neighborhood relationships without flattening the two-dimensional structure, and significantly improve the generation quality and efficiency. In handwritten trajectory modeling, the single-scale autoregressive method is difficult to take into account both global glyphs and local stroke details [25, 38]. Existing work may be done using diffusion models to approximate the trajectory distribution through multi-step denoising, or to disassemble the reconstruction into multiple stages such as skeleton type, stroke refinement, and dynamic parameter optimization [34, 14]. Inspired by VAR's "coarse-to-thin, cross-scale prediction", we rewrite the handwritten trajectory reconstruction as a multi-scale, step-by-step refinement task. The global structure is first restored on the low-scale coordinate sequence, and then the stroke details are gradually completed at the higher scale, so as to suppress the accumulation of errors while maintaining the overall consistency.

3. Method Description

3.1. Method Overview

This paper targets online handwriting scenarios: by adaptively optimizing trajectories corrupted by stroke jitter, displacement, and random unintended ligatures, it improves character legibility while emulating the user's writing style. Noting that time-indexed coordinate sequences more faithfully encode motor patterns and are inherently resolution-independent and amenable to on-device rendering, we represent handwriting trajectories as temporal coordinate sequences (cf. Sec. 3.2.2). In the absence of paired perfect–imperfect data, we simulate imperfections by injecting jitter, offsets, and connections into pen traces and apply data augmentation to increase diversity (cf. Sec. 3.2.3; Sec. 3.2.4). We propose a multi-scale autoregressive reconstruction framework for discrete trajectory sequences, formulating reconstruction as coarse-to-fine progressive refinement (cf. Sec. 3.3). The framework comprises two stages: in the content-preservation stage, the original defective trace guides generation, and explicit geometric and temporal inductive biases keep the process aligned with glyph skeletons and writing order (cf. Sec. 3.4); in the style-aggregation stage, the model learns style features from multiple reference coordinate-sequence modalities and perturbs amplitude components to control trajectory style (cf. Sec. 3.5). Finally, three loss functions minimize the discrepancy between generated and ground-truth trajectory coordinates to optimize the model (cf. Sec. 3.6).

3.2. Datasets and Data Preprocessing

3.2.1 Datasets

This study focuses on reconstructing online handwriting trajectories, whereas the existing literature has predominantly emphasized handwritten character recognition and generally lacks high-quality online handwriting datasets suitable for training and evaluation. Consequently, we em-

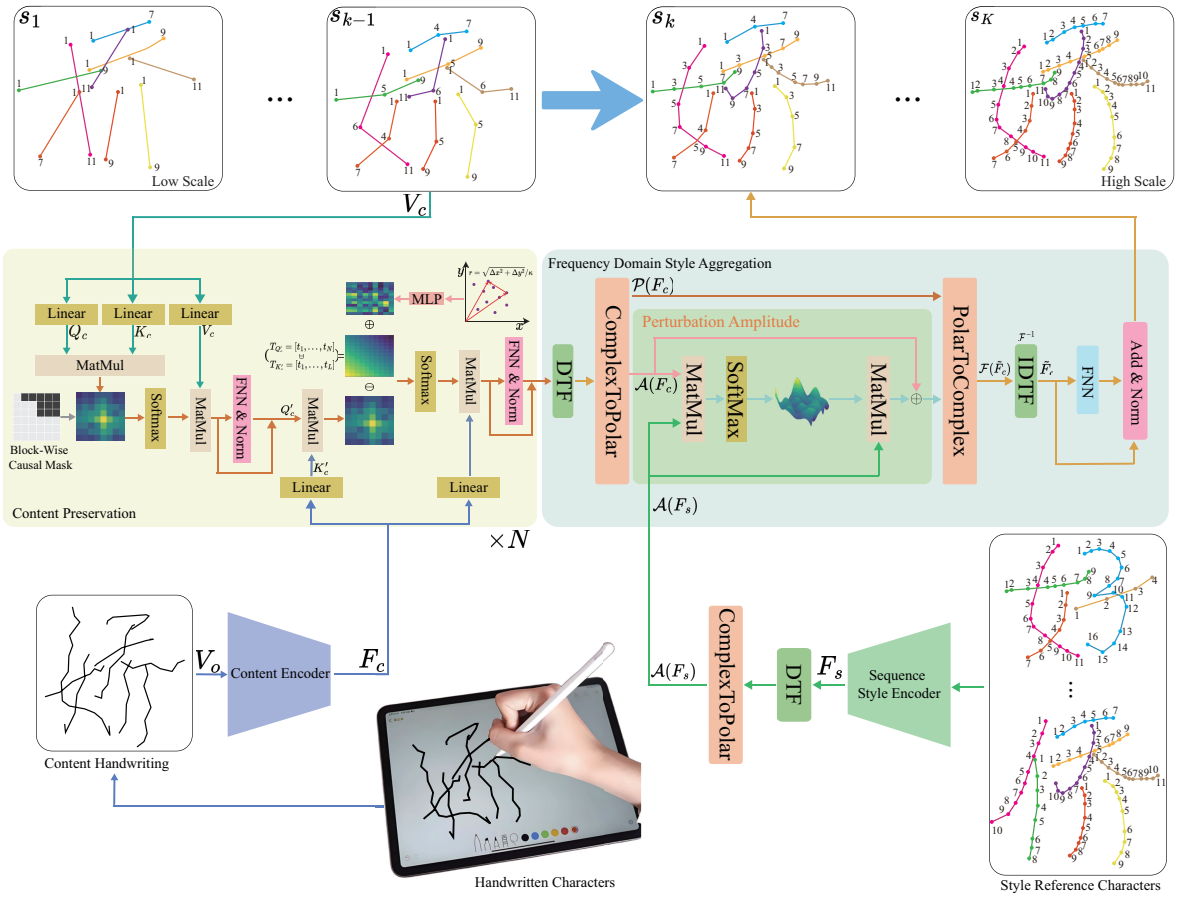

Figure 2. Overview of the proposed method. Targeting online handwriting, the model learns a user's writing style from a small set of character trajectory coordinate sequences and refines trajectories affected by stroke jitter, displacement, and random ligatures. We formulate trajectory reconstruction as a multi-scale, coarse-to-fine progressive refinement: a content-preservation stage guided by character content, followed by a style-aggregation stage that controls style by perturbing amplitude components while keeping phase components fixed.

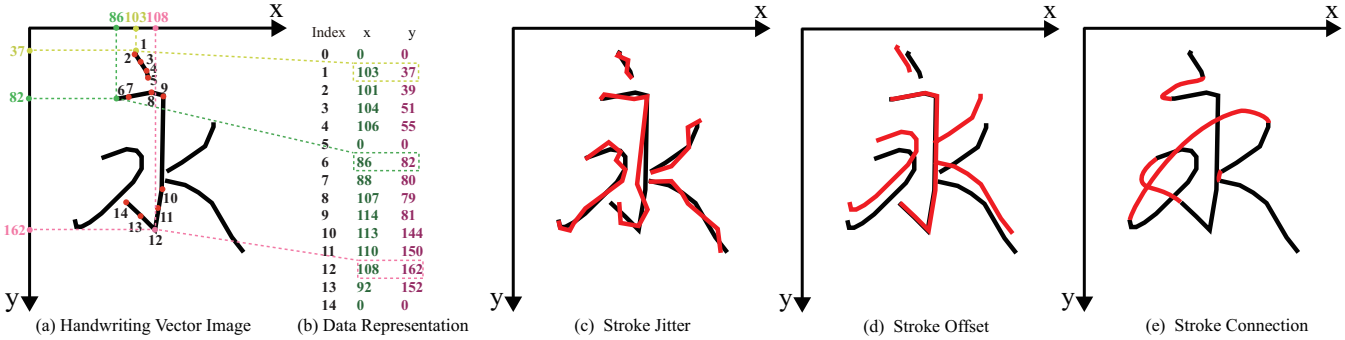

Figure 3. (a) Example of handwritten character. (b) Data representation of handwriting trajectories as coordinate sequences. (c) Simulate stroke jitter. (d) Simulate stroke layout perturbation.(e) Simulate stroke connection, where the red strokes serve as supplementary strokes for the simulated connections.

ploy the CASIA-OLHWDB (1.0–1.2) online handwritten Chinese character datasets [19]. As shown in Fig. 3(a), these datasets were collected using an Anoto digital pen and specialized dot-pattern paper, and they contain stroke start/end indicators as well as full trajectory coordinate information. The training set contains approximately 3.7 million online handwritten Chinese character samples from 1,020 writers, whereas the test set comprises 60 writers, with each writer providing 3,755 characters. In addition, we utilize the GIAHCC-UCAS2024 dataset [50], which in-

cludes 3,811 characters and a total of 368,688 samples.

3.2.2 Data Representation

We devise a structured coordinate representation for online handwriting trajectories to achieve a unified data representation across both datasets. Figure 3(b) illustrates a handwritten character as a temporally ordered sequence of N coordinate points $p_i = (x_i, y_i)$. The onset of each stroke is marked by inserting the sentinel token $(0, 0)$ to indicate the stroke boundary, and sequences shorter than N are padded with $(-1, -1)$ to ensure length alignment.

3.2.3 Handwritten Defect Dimulation

We address the scarcity of both "perfect" and "imperfect" online handwriting data by perturbing trajectory coordinates to emulate stroke jitter, stroke-level layout offsets, and stroke concatenation observed in human writing. **Stroke jitter**. We displace each coordinate in the trajectory by $\Delta_{x,y} \in [-5, 5]$ to mimic fine-grained tremor during writing (Fig. 3(c)). **Stroke offset**. The interval between adjacent $(0, 0)$ tokens in the coordinate sequence corresponds to one stroke. We apply a random displacement $\Delta_{x,y} \in [-5, 5]$ to each entire stroke to simulate layout perturbations (Fig. 3(d)). **Stroke concatenation**. We randomly connect a subset of strokes using quadratic Bézier curves to form continuous trajectories, reflecting speed-induced cursive effects (Fig. 3(e)).

3.2.4 Data Augmentation

Data augmentation uses two methods: **scaling** and **translation**.

Scaling: The lines of the handwritten track are represented using a quadratic Bézier curve formula $B(t) = (1-t)p_0 + tp_1$, where p_0 and p_1 are the Bézier curve control points. To vary the endpoint of the lines, we keep the starting control point p_0 fixed and calculate $B(t)$ instead of p_1 by randomly selecting tt from $t \in \{0.8, 0.9, 1.0, 1.1, 1.2\}$. This method produces a slightly altered writing track while retaining the main characteristics of the track.

Translation: We add a random offset Δ_x to all x-coordinates, with $\Delta_x \in [-5, 5]$, and similarly add an offset Δ_y to the y-coordinates. This method horizontally and vertically shifts the handwritten characters, thereby augmenting the data.

3.3. Next-Scale Prediction

To better align with the human perceptual principle of progressing from global structures to local details, we have transformed the autoregressive modeling of handwritten trajectory coordinates from the conventional "next-token prediction" to a "next-scale prediction" paradigm. This approach operates through a "intra-block parallel, inter-block autoregressive" mechanism, unifying global shape and fine-grained details within a unified generative framework. The coarse scale constrains the overall stroke trajectory and character layout, while the fine scale refines stroke transitions and intricate details, thereby achieving closer alignment with human perceptual patterns.

To formalize this, let the coordinate parameters of the handwritten trajectory be represented as a vector $X \in \mathbb{R}^N$. The index set $1, \ldots, N$ is partitioned into a sequence of handwritten trajectory coordinates of progressively increasing lengths:

$$I^{(1)} \subset I^{(2)} \subset \cdots \subset I^{(K)}, \quad |I^{(k)}| = n_k = \left\lfloor \frac{N}{K} k \right\rfloor. \quad (1)$$

The sequence at the k-th scale is denoted as $S^{(k)} = X_{I^{(k)}}$. The joint distribution of the complete trajectory is factorized into autoregressive components across scales as follows:

$$p(S^{(1)}, \ldots, S^{(K)}) = \prod_{k=1}^{K} p(S^{(k)} \mid S^{(1)}, \ldots, S^{(k-1)}). \quad (2)$$

Unlike the first-order autoregression that operates token-by-token, our method generates the conditional distribution for the entire coordinate sequence of length n_k at each scale in a single step. Let the newly introduced incremental block at step k be defined as:

$$C^{(k)} = I^{(k)} \setminus I^{(k-1)}, \qquad C^{(1)} = I^{(1)}. \quad (3)$$

where $I^{(k)}$ denotes the complete set of positions to be determined up to scale k.

During inference, the process proceeds sequentially across scales for $k = 1, \ldots, K$. At step k, the model conditions on the coordinate parameter sequence $\widehat{S}^{(k-1)}$ determined at the previous scale, computes the conditional distribution over all positions in set $I^{(k)}$ through a single forward pass, and generates candidate samples:

$$\tilde{S}^{(k)} \sim p_\theta\big(S^{(k)} \mid \widehat{S}^{(k-1)}\big). \quad (4)$$

The procedure then executes two operations: first, it locks the positions determined at the previous scale by keeping the coordinate parameters on $I^{(k-1)}$ unchanged; second, it writes the candidate values for the newly introduced incremental indices $C^{(k)} = I^{(k)} \setminus I^{(k-1)}$. This is formally expressed as:

$$\widehat{S}^{(k)}\big|_{I^{(k-1)}} \leftarrow \widehat{S}^{(k-1)}, \qquad \widehat{S}^{(k)}\big|_{C^{(k)}} \leftarrow \tilde{S}^{(k)}\big|_{C^{(k)}}. \quad (5)$$

Equation (5) indicates that by step k, the coordinate parameters on $I^{(k-1)}$ have already been determined and are not resampled; only the new positions in $C^{(k)}$ are populated with values obtained from Equation (4).

To supervise all K scales during training without violating causality, we introduce a block-wise causal mask in the self-attention mechanism. Let the block index of position i be denoted as:

$$b(i) = \min\{\, k : i \in I^{(k)} \,\} \in \{1, \ldots, K\}. \tag{6}$$

The attention mask matrix is defined as:

$$\mathcal{A}_{ij} = \begin{cases} 0, & b(j) \le b(i), \\ -\infty, & b(j) > b(i), \end{cases} \tag{7}$$

where, position i can only attend to contexts at its current or coarser scales. Since the mask is applied in a single operation, all scales can be supervised in parallel during training through a single forward pass, while during inference, the process unfolds sequentially in the order of $k = 1, \ldots, K$.

3.4. Content Preservation

When modeling handwriting trajectories at different scales, the over-smoothed coarse-scale trajectories tend to compromise character content features. To mitigate this issue, we introduce the drawing parameters $V_o \in \mathbb{R}^{N\times 2}$ of the original ("imperfectly written character") as content guidance. This denotes the original, imperfect, and complete handwriting trajectory coordinate sequence input by the user, consisting of N two-dimensional coordinate points. Unlike the attention modeling paradigm in natural language processing, which primarily relies on content similarity, coordinate sequences of handwriting trajectories incorporate coupled dependencies in both two-dimensional geometry and the temporal dimension. Based on this, we explicitly inject geometric and temporal inductive biases into the multi-scale autoregressive prediction, ensuring that the coarse-to-fine generation process consistently aligns with the character skeleton and writing order, thereby suppressing information loss at lower scales and enhancing cross-scale consistency.

Specifically, V_o is first transformed through feature embedding and processed by a content encoder composed of three self-attention layers, yielding content features $F_c \in \mathbb{R}^{N\times d}$. For the input sequence $S^{(k)}$ at the k-th scale, $V_c \in \mathbb{R}^{L\times 2}$ denotes the handwritten trajectory coordinate sequence at the current scale. The drawing parameters V_c at the k-th scale are embedded to obtain scale-specific content features $F_c^{(k)} \in \mathbb{R}^{L\times d}$. For intra-scale self-attention, we define:

$$Q_c = F_c^{(k)} W_Q, \quad K_c = F_c^{(k)} W_K, \quad V_c = F_c^{(k)} W_V, \tag{8}$$

The corresponding attention weights and output are given by:

$$\begin{aligned} A_c &= \mathrm{softmax}_j\Big(\tfrac{Q_c K_c^\top}{\sqrt{d}} + M^{(e)}\Big) \in \mathbb{R}^{L\times L}, \\ F_c' &= A_c\, V_c, \end{aligned} \tag{9}$$

where $\mathrm{softmax}_j$ denotes normalization along the key dimension j, and $M^{(e)}$ represents the intra-scale block-wise causal mask.

To aggregate content information across different scales, we apply cross-attention between F_c' and F_c. Let:

$$Q_c' = F_c' W_Q', \quad K_c' = F_c W_K', \quad \mathcal{F}(F_c) = F_c W_V', \tag{10}$$

and explicitly incorporate geometric and temporal relative distances into the scaled dot-product logits. Let the temporal indices of query position i and key position j be t_i and t_j, and their spatial coordinates be $\mathbf{x}_i = (x_i, y_i)$ and $\mathbf{x}_j = (x_j, y_j)$, respectively. We define:

$$\begin{aligned} \Delta t_{ij} &= t_i - t_j, & \Delta \mathbf{x}_{ij} &= \big(x_i - x_j,\ y_i - y_j\big), \\ r_{ij} &= \|\Delta \mathbf{x}_{ij}\|_2, & \tilde{r}_{ij} &= r_{ij}/\kappa, \end{aligned} \tag{11}$$

where $\kappa > 0$ is a normalization factor for coordinate scaling. For each attention head h, an additive bias is applied to the "logits", incorporating a weighted combination of geometric and temporal relative distances:

$$Z_{ij} = \frac{Q'_{c,i}\, K'^{\top}_{c,j}}{\sqrt{d}} - m_h \left|\Delta t_{ij}\right| + g_h\big(\tilde{r}_{ij}\big), \tag{12}$$

where $m_h > 0$ denotes the linear temporal slope of the h-th attention head, and the spatial bias is given by a scalar mapping function:

$$g_h\big(\tilde{r}_{ij}\big) = \mathrm{MLP}_h\big([\Delta x_{ij},\, \Delta y_{ij},\, \tilde{r}_{ij}]\big). \tag{13}$$

Based on the aforementioned attention logits, we construct the following cross-attention representation to establish a connection between the global content features and the current scale:

$$\begin{aligned} A_c' &= \mathrm{softmax}_j\big(Z + M^{(s)}\big) \in \mathbb{R}^{L\times N}, \\ \tilde{F}_c &= A_c'\, \mathcal{F}(F_c), \end{aligned} \tag{14}$$

where $M^{(s)}$ denotes the cross-scale block-wise causal mask.

3.5. Frequency Domain Style Aggregation

Under the discrete Fourier representation, let the spectrum of the handwritten trajectory signal be denoted as $\mathcal{F}(X) = \mathcal{A}(X) \odot e^{\mathrm{i}\Phi(X)}$. We posit that style information is primarily reflected in the spectral energy distribution of the magnitude $\mathcal{A}(X)$, while temporal and structural characteristics are governed by the phase $\Phi(X)$. Based on this, a style aggregation operator is designed to perturb $\mathcal{A}(X)$ while preserving $\Phi(X)$ unchanged, thereby enabling style manipulation without altering the semantic structure of the trajectory.

To rigorously analyze the separability of style-related factors and character semantic structural factors in the frequency domain, we apply the discrete Fourier transform

along the temporal dimension to the style aggregation input tensor $S_s \in \mathbb{R}^{L\times d}$, yielding:

$$\mathcal{F}(S_s) = \mathcal{A}(S_s) \odot e^{\mathrm{i}\Phi(S_s)}, \tag{15}$$

We denote the family of operators that perturb only the magnitude spectrum as:

$$\mathcal{G} = \Big\{ g \mid \mathcal{F}\big(g(S_s)\big) = \big(\mathcal{A}(S_s) + \Delta\mathcal{A}(S_s)\big) \odot e^{\mathrm{i}\Phi(S_s)}, \\ \|\Delta\mathcal{A}(S_s)\| \le \tau \Big\}, \tag{16}$$

Due to the bounded linearity of $\mathcal{F}^{-1}$, we obtain:

$$\begin{aligned} \big\|g(S_s) - S_s\big\| &= \big\|\mathcal{F}^{-1}\big((\Delta\mathcal{A}(S_s)) \odot e^{\mathrm{i}\Phi(S_s)}\big)\big\| \\ &\le \|\mathcal{F}^{-1}\|\,\|\Delta\mathcal{A}(S_s)\| \le c\,\tau. \end{aligned} \tag{17}$$

where ($\mathcal{F}^{-1}$ denotes the inverse discrete Fourier transform, $\|\cdot\|$ is the Euclidean norm for vectors and the corresponding induced operator norm for linear maps, $\Delta\mathcal{A}(S_s)$ is the perturbation applied to the magnitude spectrum, $\tau > 0$ is a prescribed upper bound on the perturbation , and $c := \|\mathcal{F}^{-1}\|$ is a constant depending only on the normalization of the transform .

Let $\mathcal{E}_\theta$ be the encoder within the style aggregation module, which outputs the feature:

$$Z \;=\; \mathcal{E}_\theta(S_s) \in \mathbb{R}^{d_z}, \qquad Z = \big[\, Z_\Phi,\; Z_\mathcal{A} \,\big], \tag{18}$$

where Z_Φ and $Z_\mathcal{A}$ represent the phase-related and magnitude-related sub-representations, respectively.

Let $u \in \mathbb{R}^{d_z}$ be the parameters of the style prediction head, s the style label, and $\mathcal{L}(\cdot, s)$ the style discrimination loss. Consider the empirical risk under the magnitude-perturbation-only operator family $\mathcal{G}$:

$$\widehat{\mathcal{R}}_{\mathrm{sty}} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{g\sim\mathcal{G}}\Big[\, \mathcal{L}\big(u^\top \mathcal{E}_\theta\big(g(S_s^{(m)})\big),\; s^{(m)}\big) \,\Big]. \tag{19}$$

Denote $Z' = \mathcal{E}_\theta\big(g(S_s)\big)$. Performing a first-order Taylor expansion at Z yields:

$$\begin{aligned} \mathcal{L}\big(u^\top Z', s\big) &\approx \mathcal{L}\big(u^\top Z, s\big) + \nabla_Z \mathcal{L}\big(u^\top Z, s\big)^\top (Z' - Z) \\ &= \mathcal{L}\big(u^\top Z, s\big) + \nabla_{Z_\Phi} \mathcal{L}^\top (Z'_\Phi - Z_\Phi) \\ &\quad + \nabla_{Z_\mathcal{A}} \mathcal{L}^\top (Z'_\mathcal{A} - Z_\mathcal{A}). \end{aligned} \tag{20}$$

Since g only modifies $\mathcal{A}(S_s)$ while keeping $\Phi(S_s)$ unchanged, the phase-related pathway of the encoder remains approximately invariant, whereas the magnitude-related pathway undergoes significant changes: there exist $0 < \delta \ll \varepsilon$ such that:

$$\|Z'_\Phi - Z_\Phi\| \;\le\; \delta, \quad \|Z'_\mathcal{A} - Z_\mathcal{A}\| \;\ge\; \varepsilon. \tag{21}$$

Substituting the above into the Taylor expansion and taking expectations, we find that the dominant term originates from the magnitude direction:

$$\mathbb{E}_{g\sim\mathcal{G}}\left[\nabla_Z \mathcal{L}^\top (Z' - Z)\right] \;\approx\; \mathbb{E}_{g\sim\mathcal{G}}\left[\nabla_{Z_\mathcal{A}} \mathcal{L}^\top \left(Z'_\mathcal{A} - Z_\mathcal{A}\right)\right]. \tag{22}$$

Under this theoretical framework, we propose a Fourier domain-based style fusion module. This module employs a cross-attention mechanism to aggregate amplitude features from multi-source style references. By applying controlled perturbations to the amplitude spectrum of the content character while preserving the phase component unchanged, it achieves style manipulation while ensuring the stability of character semantic representation. Specifically, In each iteration, for content handwritten strokes, K handwritten character sequences are randomly selected as style references. These are then passed through a style encoder, which consists of a six-layer multi-head self-attention transformer, to extract style features $F_s \in \mathbb{R}^{(K\times L)\times d}$.

To perform style aggregation and constraint in the frequency domain, we first transform the content and style features into the frequency domain and explicitly separate their magnitude (amplitude) and phase components. Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the discrete Fourier transform (DFT) and its inverse applied along the sequence dimension (of length L). Suppose the features obtained from the content branch are $F_c \in \mathbb{R}^{L\times d}$, and the output of the style encoder is $F_s \in \mathbb{R}^{(K\times L)\times d}$. Then:

$$\mathcal{F}(\tilde{F}_c) \in \mathbb{C}^{L\times d}, \quad \mathcal{F}(F_s) \in \mathbb{C}^{(K\times L)\times d}. \tag{23}$$

From this, the magnitude spectra and phase spectra of the content and style can be derived as:

$$\mathcal{A}(F_c) = |\mathcal{F}(F_c)|, \quad \Phi(F_c) = \angle\mathcal{F}(F_c), \quad \mathcal{A}(F_s) = |\mathcal{F}(F_s)|, \tag{24}$$

where $|\cdot|$ denotes the complex modulus and $\angle(\cdot)$ represents the phase. We then use the content magnitude $\mathcal{A}(F_c)$ as the query, and the set of style magnitudes $\mathcal{A}(F_s)$ as keys and values, to aggregate amplitude information from the K references through cross-attention:

$$\hat{\mathcal{A}}(F_s) = \mathrm{Attn}\big(\mathcal{A}(F_c)W_Q,\, \mathcal{A}(F_s)W_K,\, \mathcal{A}(F_s)W_V\big), \tag{25}$$

where W_Q, W_K, W_V are learnable projection matrices. To achieve controlled perturbation, we employ residual-style magnitude modulation:

$$\tilde{\mathcal{A}}(F_c) = \mathcal{A}(F_c) + \lambda\,\sigma\big(\hat{\mathcal{A}}(F_s) - \mathcal{A}(F_c)\big), \tag{26}$$

where $\lambda \in [0, 1]$ is the style intensity coefficient and $\sigma(\cdot)$ denotes the element-wise $\tanh$ gating function. The modulated magnitude is then combined with the original phase to reconstruct the frequency-domain features, which are subsequently transformed back to the time domain via the inverse transform:

$$\mathcal{F}(\tilde{F}_c) = \tilde{\mathcal{A}}(F_c) \odot e^{\,j\Phi(F_c)}, \quad \tilde{F}_c = \mathcal{F}^{-1}\big(\mathcal{F}(\tilde{F}_c)\big). \tag{27}$$

The above process ensures that only the amplitude spectrum of the content character undergoes constrained perturbation to incorporate stylistic features, while the phase component $\Phi(F_c)$ is entirely preserved, thus maintaining the semantic integrity of the handwriting trajectory.

3.6. Loss Function

During training, we optimize the model by minimizing the discrepancy between the model-generated handwritten trajectory coordinates and the ground-truth target trajectory coordinates. Using the mean squared error (MSE) as the loss function, the loss for the generated coordinate sequence $\hat{Y} \in \mathbb{R}^2$ and the target coordinate sequence $Y \in \mathbb{R}^2$ is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{L}\sum_{i=1}^{L} \|y_i - \hat{y}_i\|^2. \tag{28}$$

In addition, we simultaneously impose a style-consistency constraint in the frequency domain. Let $\mathcal{F}_r = E_S(S_r)$, $\mathcal{F}_g = E_S(\hat{Y})$, and $\mathcal{F}_t = E_S(Y)$ denote the style features extracted by the style encoder. We design the following two loss terms:

$$\begin{cases} \mathcal{L}_{\text{rg}} & = \left\|\mathcal{A}(\mathcal{F}_r) - \mathcal{A}(\mathcal{F}_g)\right\|^2, \\ \mathcal{L}_{\text{rt}} & = \left\|\mathcal{A}(\mathcal{F}_r) - \mathcal{A}(\mathcal{F}_t)\right\|^2. \end{cases} \tag{29}$$

4. Experiments

4.1. Implementation Details

We train the model using the AdamW optimizer, setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to 2×10^{-4}. The model is trained for 500 epochs on 8 RTX 4090 GPUs. Each training iteration randomly samples one character from CASIA-OLHWDB [19] or GIAHCC-UCAS2024 [50] as the input character, and four characters from CASIA-OLHWDB [19] as the target and style-reference characters.

4.2. Evaluation Metrics

DTW: Dynamic Time Warping (DTW) [5] is used to compute the distance between two sequences of different lengths. Therefore, we use DTW to evaluate the similarity between real and generated handwriting, where a lower DTW value indicates higher similarity.

Content Score: We render handwritten characters from the CASIA-OLHWDB dataset as 256×256 images; then train the pre-trained ResNet-50 using the Adam optimizer with a learning rate of 0.001. The ground-truth character class probability is used as the content consistency score, averaged over 1000 characters.

Style Score: For the style score, we train a ResNet-50 network on the test set to recognize which writer produced each character. The writer-identification probability serves as the style consistency score, averaged over 1000 characters.

User Preference: We invited 40 volunteers to evaluate 200 characters generated by each method. Characters deemed indistinguishable from the target font are counted as "correct", and the average number of correct characters is used as the quantitative metric.

4.3. Comparison with Other Methods

Our proposed method is compared with handwriting font generation approaches, including Drawing WriteLikeYou [32], Diff-Writer [30], ElegantlyWritten [25], and EHW-Font[38] which model handwriting traces as sequential data. DeepImitator [49] and SDT [10] transform handwriting traces into images to generate stylized handwriting trace sequences.

4.3.1 Quantitative Comparison

Table 1. Quantitative comparison of different methods under the USSC and USUC settings on the Chinese dataset.

Setting	Method	Style ↑	Content ↑	DTW ↓	User(%) ↑
USSC	Diff-Writer [30]	41.02	72.23	1.4852	12.2
	DeepImitator[49]	52.88	87.18	1.3011	16.6
	WriteLikeYou[32]	76.21	94.25	1.1101	47.5
	SDT[10]	85.31	95.71	0.9554	55.2
	ElegantlyWritten [25]	89.10	96.04	0.8135	59.1
	EHW-Font[38]	91.85	96.43	0.8012	61.2
	Ours	95.65	96.86	0.7822	64.6
USUC	Diff-Writer [30]	33.42	64.33	2.1037	8.1
	DeepImitator[49]	40.24	71.79	1.8123	10.4
	WriteLikeYou[32]	64.21	90.24	1.3923	37.5
	SDT[10]	80.31	93.51	1.1954	50.2
	ElegantlyWritten [25]	85.10	94.12	0.8978	56.3
	EHW-Font[38]	86.33	94.23	0.8643	57.1
	Ours	91.37	95.76	0.8023	61.3

To comprehensively evaluate each method, we consider two testing scenarios: Unseen Writing Style with Seen Characters (USSC) and Unseen Writing Style with Unseen Characters (USUC). USSC covers the 2,000 characters that appear during training, whereas USUC consists solely of characters absent from the training set.

Tab. 1 presents a quantitative comparison between our method and six representative baselines. Overall, our method achieves state-of-the-art performance across multiple evaluation metrics. Notably, DeepImitator [49], WriteLikeYou [32], and SDT [10] jointly model style features from multiple reference images, and EHW-Font [38] further extends multi-character style learning to both image and sequence modalities. Nevertheless, whether one can explicitly model and leverage multi-granularity style cues across multiple references remains a key factor affecting performance. Our method controls style by applying controlled perturbations to the amplitude spectrum of the content characters while keeping the phase component unchanged, thereby

preserving stable semantic representations of the characters. Another general observation is that content-related scores are typically higher than style-related scores, indicating that, compared with legibility restoration, achieving realistic style imitation is more challenging.

Under the more stringent USUC setting , all methods except ours exhibit a pronounced drop in performance, highlighting the superior generalization of our approach to unseen characters. Unlike typical autoregressive or non-autoregressive paradigms, we frame handwritten trajectory reconstruction as a multi-scale, coarse-to-fine hierarchical refinement process: we first recover the global character structure and stroke layout at a coarse level, and then progressively inject stroke trajectories and local details. This design significantly surpasses prior state-of-the-art methods on style imitation metrics while maintaining high fidelity in both global structure and fine-grained stroke rendering.

4.3.2 Qualitative Comparison

As shown in Fig. 4, we compare several approaches to handwritten trajectory reconstruction. DeepImitator [49] combines a Gaussian mixture model with conditional GRUs and aggregates complementary style cues from multiple reference images, yielding stronger reproduction of complex styles. Building on this trend, SDT [10] and Elegantly-Written [25] adopt Transformer-based architectures with multiple style references, where self-attention helps capture long-range dependencies across strokes and time. Further, EHW-Font [38] extends ElegantlyWritten with dual-modality learning, leading to more stable style representations. Distinct from the aforementioned autoregressive or non-autoregressive paradigms, our approach adopts a progressive resolution evolution strategy. By prioritizing the global layout at coarser scales before refining local details, we effectively mitigate the structural distortion and over-smoothing often observed in baseline methods. This hierarchical dependency ensures that local stroke nuances are generated within a stable global context, yielding reconstructions that are structurally complete while remaining stylistically consistent.

4.4. Ablation Study

4.4.1 Effectiveness of the Content Guidance Module

To further validate the design of the content guidance module, we conduct ablation experiments under multiple configurations, as shown in Tab. 2. The baseline model removes the entire content guidance module and relies solely on the decoder's self-attention for modeling trajectory sequences. This setting results in notably lower Style and Content scores, indicating insufficient alignment between the generated trajectory and the underlying character structure.

Introducing a standard cross-attention mechanism significantly improves performance, suggesting that modeling global content-to-sequence alignment is beneficial. However, without structural inductive biases, spatial and temporal misalignment still occurs. To address this, we separately incorporate temporal and geometric biases into the attention logits. Experimental results show that the geometric bias contributes more significantly than the temporal counterpart, highlighting the importance of spatial configuration in preserving the overall character shape. The complete content guidance module, which integrates both temporal and geometric biases, achieves the best performance across all metrics. It consistently improves style consistency, content accuracy, and trajectory alignment (as measured by DTW), confirming the effectiveness and complementarity of the proposed biases.

Model Variant	Style ↑	Content ↑	DTW ↓
w/o CG	83.33	89.46	1.0821
CA (no Bias)	90.11	93.58	0.8465
CA + Temp	91.35	94.02	0.8234
CA + Geo	93.18	95.17	0.7971
CA + Full (Ours)	**95.65**	**96.86**	**0.7822**

Table 2. Ablation study on the content guidance module. "w/o CG": without content guidance; "CA": cross-attention; "Temp": temporal bias; "Geo": geometric bias. The full bias configuration achieves the best performance in style preservation, content reconstruction, and spatiotemporal alignment.

4.4.2 Effectiveness of Style Aggregation

To evaluate the effectiveness of the individual components in our frequency-domain style aggregation module, we conducted a series of ablation experiments. The results are presented in Tab. 3. We first replaced the frequency-domain design with a time-domain baseline, where standard cross-attention is directly applied to the original representations F_c andF_s in the spatial domain. This configuration neglects spectral structure and performs poorly in both style consistency (Style) and content preservation (Content), demonstrating the importance of frequency-domain modeling for style representation. Next, we performed cross-attention in the frequency domain but perturbed both magnitude and phase components simultaneously. Although this setting enhances stylistic variation, it disrupts the semantic structure of handwritten characters, leading to a significant increase in DTW, thereby highlighting the necessity of preserving phase stability. We then separately perturbed the magnitude and phase components to examine their respective roles. Results show that perturbing only the magnitude yields the best performance, suggesting that amplitude primarily governs stylistic expression, while phase governs

Figure 4. A qualitative comparison with the state-of-the-art online Chinese trace generation methods is presented.

Model Variant	Style ↑	Content ↑	DTW ↓
Time-domain Cross-Attn	88.42	91.07	0.9448
Freq Cross-Attn (Mag + Phase)	91.36	93.15	0.8725
Freq Attn + Phase Perturb.	89.91	92.33	0.9024
Freq Attn + Mag Perturb. (Ours)	**95.65**	**96.86**	**0.7822**

Table 3. Ablation study on frequency-domain style aggregation. Only perturbing amplitude while keeping phase intact (Ours) yields the best trade-off between stylistic richness and structural fidelity.

structural integrity. In contrast, perturbing only the phase not only degrades Style and Content scores but also introduces visible distortions to trajectory structure. Finally, our full method (Ours), which perturbs only the magnitude spectrum while keeping the phase unchanged, achieves the best balance between stylistic enhancement and semantic fidelity.

4.4.3 Effectiveness of Multi-Scale Modeling

To assess the effect of the multi-scale autoregressive strategy on handwriting reconstruction, we evaluate six scale partitions on the USUC dataset under the same model architecture and training hyperparameters. When the scale is set to 1, the model degenerates into a single-step, non-autoregressive prediction that outputs all parameters at once. As shown in Fig. 5, the style and content scores are the lowest and the DTW is the highest at scale 1; increasing the scale from 1 to 4 yields a "leap" in performance and constitutes the main gain region. Further increasing the scale to 7 brings small additional improvements in Style and Content, while DTW decreases further, reaching a global optimum around 8; when expanded to 14, the quality gain saturates and slightly regresses. Meanwhile, the inference latency grows approximately linearly with the number of scales.

Multi-scale autoregression first captures global geometry and layout and then refines local details at finer scales; therefore it markedly reduces early reconstruction error compared with single-step prediction. However, when the granularity becomes overly fine, the marginal benefit quickly diminishes due to limited model capacity and data diversity, and the longer prediction chain introduces error accumulation and higher inference cost. Balancing quality and efficiency, we adopt scale 8 as the default in subsequent experiments. With larger models or richer data, the optimal scale may shift to the right.

4.5. Analysis

4.5.1 Real Handwritten Character Optimization

To assess the effectiveness of the proposed strategy in real-world use, we established an evaluation setup that mirrors the acquisition conditions of the source dataset, employing

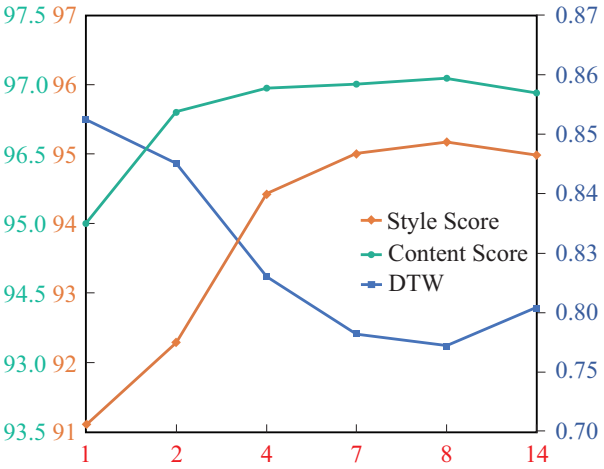

Figure 5. Quantitative evaluation of character reconstruction under different numbers of scales.

an Anoto digital pen and custom dot-pattern paper. We recruited 10 writers, each of whom produced two sets of data: (i) 200 deliberately perturbed, jitter-prone non-ideal characters and (ii) 200 carefully crafted ideal characters. These data constitute our real-handwriting test set. In parallel, we randomly sampled 200 characters from CASIA-OLHWDB and generated corresponding degraded versions following the proposed simulation pipeline, yielding a simulated evaluation set for comparison against the real data.

Quantitative results are reported in Tab. 4. Reconstruction metrics on the real-captured data are slightly superior to those on the simulated data. This observation indicates that the writing styles and degradation patterns learned from the zero-label simulated samples transfer robustly to genuine handwriting trajectories, enabling high-fidelity recovery of both content and style in real scenarios. Qualitative analyses are presented in Fig. 6. The reconstructed trajectories align closely with their ideal references in terms of stroke onsets and offsets, stroke order and direction, and the geometry of key inflection points. These visual results further substantiate that the label-free simulation scheme not only accurately reproduces degraded writing in the synthetic setting, but also transfers seamlessly to real capture conditions, delivering stylistically consistent and content-faithful reconstructions of real-world handwriting.

Modal	Style ↑	Content ↑	DTW ↓
$\mathcal{S}$	95.65	96.86	0.7822
$\mathcal{R}$	96.87	97.32	0.7697

Table 4. Quantitative evaluation of real handwritten character reconstruction. $\mathcal{R}$ denote real handwritten, and $\mathcal{S}$ denote simulated handwriting.

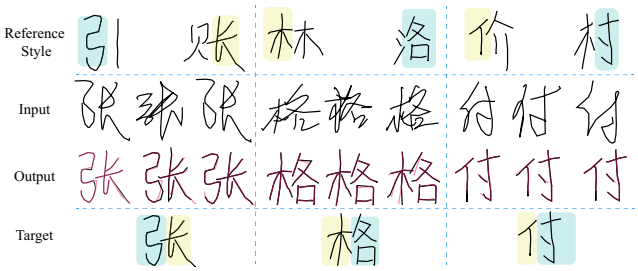

Figure 6. Qualitative evaluation of real handwritten character reconstruction.

4.5.2 Pseudo-Character Experiments.

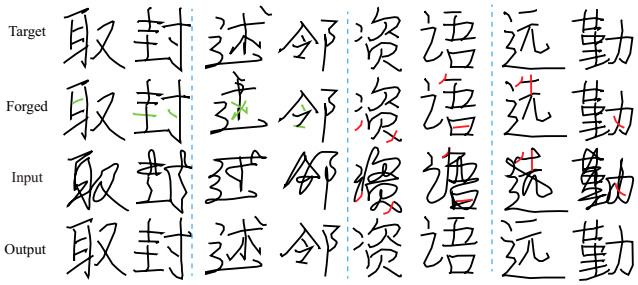

Figure 7. Pseudo-character generation experiment. Red indicates added strokes; green indicates removed strokes.

To evaluate the model's robustness to stroke-level perturbations, we construct "pseudo-characters" from test glyphs. While preserving the writer's style, we perform controlled edits by randomly or manually adding or removing one to two strokes or stroke segments, thereby creating compositions unseen in the training set. The edited samples serve as inputs and are paired with the corresponding target glyphs ; the model then reconstructs the characters to produce outputs. In Fig. 7, added strokes are marked in red and removed strokes in green. As shown in Fig. 7, the model effectively suppresses extraneous strokes and completes missing components, yielding outputs that are broadly consistent with the targets in overall structure and stroke style.

4.5.3 Reconstruction Visualization at Multiple Scales.

We conducted a systematic evaluation of the model's reconstruction performance on Chinese characters written by four individuals across scales 1–8 and provided visualized results. As shown in Fig. 8, the experiments verify that the multi-scale prediction mechanism progressively accumulates and refines fine-grained details as the levels advance. Because each level only needs to process the details newly introduced at that scale, the model achieves improved reconstruction quality with greater stability and efficiency.

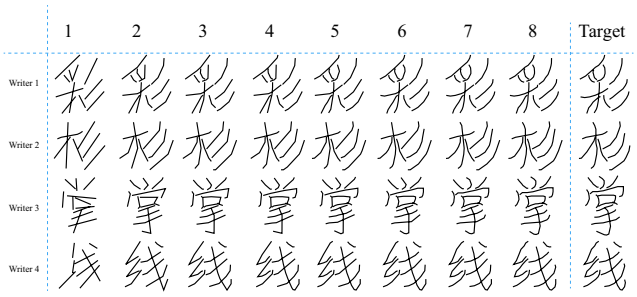

Figure 8. Visualization of reconstructed handwritten characters at different scales.

Methods	Input	Acc. ↑	Input	Acc.↑
1D-CNN	$\mathcal{O}$	68.76	$\mathcal{R}$	90.12
DCNN	$\mathcal{O}$	71.67	$\mathcal{R}$	94.37
1D-TCRN	$\mathcal{O}$	75.45	$\mathcal{R}$	95.22
PyGT	$\mathcal{O}$	80.37	$\mathcal{R}$	97.76

Table 5. Handwritten character recognition results. $\mathcal{O}$ represents distorted characters, while $\mathcal{R}$ indicates characters reconstructed using the proposed method.

4.5.4 Character Recognition Evaluation

To quantify the downstream benefit of our stroke reconstruction, we compare character recognition accuracy under two input conditions: distorted characters ($\mathcal{O}$) and characters reconstructed by our method ($\mathcal{R}$). We evaluate four competitive recognizers 1D-CNN, DCNN, 1D-TCRN, and PyGT using their standard configurations; only the input source varies between $\mathcal{O}$ and $\mathcal{R}$. As summarized in Tab. 5, averaged over all models, accuracy increases from 74.06% ($\mathcal{O}$) to 94.37% ($\mathcal{R}$). These consistent gains indicate that our reconstructed strokes better preserve discriminative handwriting patterns and reduce distortions that hinder downstream recognizers, thereby providing an indirect but informative measure of reconstruction quality.

4.5.5 Inference Efficiency Analysis

To evaluate the practical deployability of the proposed method, we compared the inference latency of our approach against representative baseline methods. All tests were conducted on a single NVIDIA RTX 4090 GPU with a batch size of 1. The average inference time per character (in milliseconds) was calculated over 1,000 runs.

Tab. 6 summarizes the experimental results. The sequence-based diffusion model, Diff-Writer [30], incurs the highest computational overhead due to its slow iterative denoising process. Among sequence-based autoregressive methods, WriteLikeYou [32], DeepImitator [49], and SDT [10] employ the standard 'next-token" prediction paradigm. Their inference time grows linearly with the sequence length, creating a bottleneck in the generation process. In contrast, we reformulate the generation process as a 'next-scale" prediction task. By generating trajectory coordinates in a coarse-to-fine manner across only 8 scales, we achieve block-wise parallel generation within each scale; this mechanism reduces the number of autoregressive steps from to 8. Furthermore, by caching the style reference character features and bypassing the style encoder, the inference time is further reduced to 386 ms, fully demonstrating its superiority for real-time applications.

Table 6. Comparison of inference speed on an RTX 4090 GPU. The proposed method significantly outperforms baselines due to its multi-scale parallel generation mechanism.

Method	Modality	Architecture	Time (ms) ↓
Diff-Writer [30]	Sequence	Diffusion	820.50
DeepImitator [49]	Sequence	GRU&CNN	495.20
SDT [10]	Sequence	Transformer&CNN	488.15
WriteLikeYou [32]	Sequence	RNN	442.30
ElegantlyWritten [25]	Sequence	Transformer	358.60
EHW-Font [38]	Seq + Img	Transformer&CNN	665.12
Ours	**Sequence**	**Transformer**	**488.42**
Ours (Style Cache)	**Sequence**	**Transformer**	**386.21**

4.5.6 Applications to Other Languages

To investigate the generalization capability of our model on unseen scripts, we conducted handwritten character reconstruction experiments on Japanese, Korean, and English. It is important to emphasize that these experiments were implemented in a zero-shot manner, meaning the model was not trained on these specific languages. As shown in Fig. 9, the proposed method effectively reconstructs these characters. Notably, compared to Chinese characters, Korean and Japanese generally feature fewer strokes and simpler topological structures, making them relatively easier for the model to process.

Figure 9. Applicability to Korean, Japanese and English handwriting

5. Conclusion

This paper proposes a method that enhances the legibility of handwritten characters while preserving the user's writing style. The model learns stylistic and dynamic handwriting patterns in the discrete coordinate-sequence modality and reformulates trajectory reconstruction as a multi-scale, coarse-to-fine refinement task. In the content preservation

stage, we inject geometric and temporal inductive biases to ensure alignment with the underlying character skeleton and stroke order throughout the generation process. In the style aggregation stage, we achieve controllable style modulation by adopting a magnitude-perturbation and phase-freezing separation strategy. Encouraging experimental results validate the effectiveness of our proposed approach. **Limitation:** The model is trained solely on a simulated dataset and does not address real-time deployment challenges. **Negative Impact:** This technique could potentially be misused for forging or mimicking individuals' handwriting or signatures.

Acknowledgement

This study was supported in part by Liaoning Provincial Science and Technology Plan Joint Program (Technology R&D Program Project) under Grants 2024JH2/102600108, the Science and Technology Innovation Fundation of Dalian under Grant 2023JJ12GX026, and in part by the Foundation of Key Laboratory of Education Informatization for Nationalities (Yunnan Normal University, Ministry of Education.) under Grant EIN2024B002.

References

[1] E. Aksan, T. Deselaers, A. Tagliasacchi, and O. Hilliges. Cose: Compositional stroke embeddings. *Advances in Neural Information Processing Systems*, 33:10041–10052, 2020. 4

[2] X. Ao, X. Li, X. Zhang, and C. Liu. Bayesian classifier calibration based on synthesized samples for zero-shot chinese character recognition. *Pattern Recognition*, page 112251, 2025. 1

[3] Y. Bai, Y. Guo, J. Wei, L. Lu, R. Wang, and Y. Wang. Fake generated painting detection via frequency analysis. In *Proceedings of the 2020 IEEE International Conference on Image Processing*, pages 1256–1260. IEEE, 2020. 2, 3

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003. 4

[5] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, pages 359–370, 1994. 9

[6] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13930–13940, 2021. 2, 3

[7] A. Carlier, M. Danelljan, A. Alahi, and R. Timofte. Deepsvg: A hierarchical generative network for vector graphics animation. *Advances in Neural Information Processing Systems*, 33:16351–16361, 2020. 4

[8] Y. Chen, H. Zhang, and C.-L. Liu. Improved learning for online handwritten chinese text recognition with convolutional prototype network. In *Document Analysis and Recognition*, pages 38–53. Springer, 2023. 1

[9] Y. Chen, H. Zhang, M. Ren, and C. Liu. Recognition of online handwritten chinese texts in any writing direction via stroke classification based over-segmentation. In *International Conference on Pattern Recognition*, pages 375–391. Springer, 2024. 1

[10] G. Dai, Y. Zhang, Q. Wang, Q. Du, Z. Yu, Z. Liu, and S. Huang. Disentangling writer and character styles for handwriting generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5977–5986, 2023. 2, 9, 10, 13

[11] M. Diaz, A. Mendoza-García, M. A. Ferrer, and R. Sabourin. A survey of handwriting synthesis from 2019 to 2024: A comprehensive review. *Pattern Recognition*, page 111357, 2025. 2

[12] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[13] J. Gan and W. Wang. Higan: Handwriting imitation conditioned on arbitrary-length texts and disentangled styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7484–7492. AAAI Press, 2021. 3

[14] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025. 2, 4

[15] J. Henry, T. Natalie, and D. Madsen. Pix2pix gan for image-to-image translation. *Research Gate Publication*, pages 1–5, 2021. 3

[16] H. Huang, D. Yang, G. Dai, Z. Han, Y. Wang, K.-M. Lam, F. Yang, S. Huang, Y. Liu, and M. He. Agtgan: Unpaired image translation for photographic ancient character generation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5456–5467, 2022. 4

[17] Z. Huang, X. Qiu, Y. Ma, Y. Zhou, C. Zhang, and X. Li. Nfig: Autoregressive image generation with next-frequency prediction. *arXiv preprint arXiv:2503.07076*, 2025. 4

[18] S.-K. Lee and J.-H. Kim. Air-text: Air-writing and recognition system. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1267–1274, 2021. 3

[19] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. Casia online and offline chinese handwriting databases. In *2011 international conference on document analysis and recognition*, pages 37–41. IEEE, 2011. 5, 9

[20] Y. Liu, F. binti Khalid, M. R. binti Mustaffa, and A. bin Azman. Dual-modality learning and transformer-based approach for high-quality vector font generation. *Expert Systems with Applications*, 240:122405, 2024. 2, 3

[21] Y. Liu, F. binti Khalid, C. Wang, M. R. binti Mustaffa, and A. bin Azman. An end-to-end chinese font generation network with stroke semantics and deformable attention skip-connection. *Expert Systems with Applications*, 237:121407, 2024. 1, 3

[22] Y. Liu, Y. Ding, F. B. Khalid, C. Wang, and L. Wang. Few-shot font generation via denoising diffusion and component-level fine-grained style. *Expert Systems with Applications*, 296:128987, 2026. 1, 3

[23] Y. Liu, Y. Ding, X. Li, b. A. Azreen, et al. Unsupervised font generation network integrating content and style representation. *Journal of Computer-Aided Design & Computer Graphics*, 37(5):865–876, 2025. 1, 3

[24] Y. Liu, F. B. Khalid, C. Wang, M. R. B. Mustaffa, and A. B. Azman. Diffvecfont: Fusing dual-mode reconstruction vector fonts via masked diffusion transformers. In *International Conference on Computational Visual Media*, pages 339–363. Springer, 2025. 2

[25] Y. Liu, F. B. Khalid, L. Wang, Y. Zhang, and C. Wang. Elegantly written: Disentangling writer and character styles for enhancing online chinese handwriting. In *European Conference on Computer Vision*, pages 409–425. Springer, 2024. 2, 3, 4, 9, 10, 13

[26] I. Memon, M. A. U. Hassan, and J. Choi. Patch-font: Enhancing few-shot font generation with patch-based attention and multitask encoding. *Applied Sciences*, 15(3):1654, 2025. 3

[27] W. Pan, A. Zhu, X. Zhou, B. K. Iwana, and S. Li. Few shot font generation via transferring similarity guided global style and quantization local style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19506–19516, 2023. 3

[28] S. Park, S. Chun, J. Cha, B. Lee, and H. Shim. Multiple heads are better than one: Few-shot font generation with multiple localized experts. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13900–13909, 2021. 3

[29] V. Pippi, S. Cascianelli, and R. Cucchiara. Handwritten text generation from visual archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22458–22467, 2023. 3

[30] M.-S. Ren, Y.-M. Zhang, Q.-F. Wang, F. Yin, and C.-L. Liu. Diff-writer: a diffusion model-based stylized online handwritten chinese character generator. In *International Conference on Neural Information Processing*, pages 86–100. Springer, 2023. 4, 9, 13

[31] L. S. F. Ribeiro, T. Bui, J. Collomosse, and M. Ponti. Sketchformer: Transformer-based representation for sketched structure. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14153–14162, 2020. 4

[32] S. Tang and Z. Lian. Write like you: Synthesizing your cursive online chinese handwriting via metric-based meta learning. *Computer Graphics Forum*, 40(2):141–151, 2021. 9, 13

[33] S. Tang, Z. Xia, Z. Lian, Y. Tang, and J. Xiao. Fontrnn: Generating large-scale chinese fonts via recurrent neural network. *Computer Graphics Forum*, 38(7):567–577, 2019. 3

[34] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems*, 37:84839–84865, 2024. 2, 4

[35] C. Wang, Y. Ding, Y. Liu, G. Zhan, and Z. Li. Chinese font generation from stroke semantic and attention mechanism. *Journal of Computer-Aided Design & Computer Graphics*, 34(8):1229–1237, 2022. 3

[36] C. Wang, M. Zhou, T. Ge, Y. Jiang, H. Bao, and W. Xu. Cf-font: Content fusion for few-shot font generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1858–1867, 2023. 3

[37] L. Wang, Y. Liu, M. Y. Sharum, R. Yaakob, K. A. Kasmiran, and C. Wang. Deep learning for chinese font generation: A survey. *Expert Systems with Applications*, 276:127105, 2025. 2, 3

[38] L. Wang, C. Wang, and Y. Liu. Ehw-font: A handwriting enhancement approach mimicking human writing processes. *Expert Systems with Applications*, 278:127278, 2025. 2, 4, 9, 10, 13

[39] Y. Xie, X. Chen, L. Sun, and Y. Lu. Dg-font: Deformable generative networks for unsupervised font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 735–751. IEEE, 2021. 3

[40] Q. Xu, R. Zhang, Z. Fan, Y. Wang, Y.-Y. Wu, and Y. Zhang. Fourier-based augmentation with applications to domain generalization. *Pattern Recognition*, 139:109474, 2023. 2, 3

[41] M. Yao, Y. Zhang, X. Lin, X. Li, and W. Zuo. Vq-font: Few-shot font generation with structure-aware enhancement and quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16407–16415, 2024. 3

[42] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 4

[43] M.-M. Yu, H. Zhang, F. Yin, and C.-L. Liu. An approach for handwritten chinese text recognition unifying character segmentation and recognition. *Pattern Recognition*, 151:110373, 2024. 1

[44] J. Zdenek and H. Nakayama. Jokergan: memory-efficient model for handwritten text generation with text line awareness. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5655–5663, 2021. 3

[45] J. Zeng, Q. Chen, Y. Liu, M. Wang, and Y. Yao. Strokegan: Reducing mode collapse in chinese font generation via stroke encoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3270–3277, 2021. 3

[46] S. Zeng and Z. Pan. An unsupervised font style transfer model based on generative adversarial networks. *Multimedia Tools and Applications*, 81(4):5305–5324, 2022. 3

[47] Y. Zhang, J. Man, and P. Sun. Mf-net: a novel few-shot stylized multilingual font generation method. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2088–2096, 2022. 3

[48] Y. Zhang, Y. Tian, and J. Hou. Csast: Content self-supervised and style contrastive learning for arbitrary style transfer. *Neural Networks*, 164:146–155, 2023. 3

[49] B. Zhao, J. Tao, M. Yang, Z. Tian, C. Fan, and Y. Bai. Deep imitator: Handwriting calligraphy imitation via deep atten-

tion networks. *Pattern Recognition*, 104:107080, 2020. 9, 10, 13

[50] W. Zhao, M. Wu, and W. Wang. Diffchar: A fast conditional diffusion model for air-writing chinese character generation. *Pattern Recognition*, page 112307, 2025. 5, 9

[51] A. Zhu, X. Lu, X. Bai, S. Uchida, B. K. Iwana, and S. Xiong. Few-shot text style transfer via deep feature similarity. *IEEE Transactions on Image Processing*, 29:6932–6946, 2020. 3