

Article

Skeleton-Aware Deformable Alignment for Few-Shot Font Generation

Songshui Wu¹, Guangyong Zheng^{1,*}, Tao Jiang² and Jinke Yang³

¹ School of Computer Science and Technology, Hengyang Normal University, Hengyang 421002, China; songshuiwu66@gmail.com

² Hunan Provincial Key Laboratory of Intelligent Information Processing and Application, Hengyang Normal University, Hengyang 421002, China; 13034837085@163.com

³ Hunan Engineering Research Center of Cyberspace Security Technology and Applications, Hengyang Normal University, Hengyang 421002, China; yjk1449134559@gmail.com

* Correspondence: hysyjsjxzy@hynu.edu.cn; Tel.: +86-139-7471-7931

Abstract

Few-shot font generation can be viewed as a challenging conditional image generation task, where the goal is to synthesize target glyphs from only a few reference samples while preserving structural fidelity and style consistency. This problem becomes particularly difficult for characters with complex spatial layouts and fine-grained stroke topology, where existing methods often struggle to simultaneously maintain structural integrity, local continuity, and stylistic coherence under sparse-reference conditions. To address this issue, we propose a skeleton-aware deformable alignment framework for few-shot font generation. Specifically, explicit skeleton priors are introduced into the diffusion-based generation process to provide structural supervision during denoising. In addition, a structure-constrained deformable content alignment module is designed to improve local feature correspondence while suppressing unreasonable geometric deformation. We further develop a multi-module content aggregation strategy to jointly model global layout patterns and local stroke details through complementary multi-level representations. Extensive experiments demonstrate that the proposed method consistently outperforms state-of-the-art approaches in both quantitative and qualitative evaluations. The results show that our method provides stronger structural preservation, better perceptual quality, and improved generalization in structurally complex glyph generation and cross-lingual style transfer.

Keywords: few-shot font generation; diffusion model; deformable content alignment; multi-module content aggregation



Academic Editors: Abu Saleh Musa Miah, Jungpil Shin and Md. Al Mehedi Hasan

Received: 27 May 2026

Revised: 25 June 2026

Accepted: 25 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Few-shot font generation (FFG) [1] aims to synthesize style-consistent target glyphs from a limited set of reference characters while preserving the content structure of unseen characters. This problem is of considerable practical importance for font design automation and low-resource font library construction. However, it remains highly challenging for complex writing systems such as Chinese, whose characters are characterized by intricate stroke configurations, strong spatial dependencies, and substantial structural diversity. Under such circumstances, achieving both global style consistency and fine-grained structural faithfulness with only a few reference samples is still an open challenge.

Existing few-shot font generation methods are predominantly built upon the paradigm of content–style disentanglement and can generally be categorized into two groups. One

group focuses on learning implicit glyph priors through deep neural networks, aiming to decouple content and style representations for subsequent style transfer. While these methods are usually effective in capturing coarse style patterns, their lack of explicit structural constraints often leads to stroke fracture and local structural distortion when handling complex characters. The other group attempts to incorporate explicit prior knowledge, such as strokes, radicals, or edge information, to enhance structural modeling. Despite their improved ability to preserve local details, these methods commonly rely on additional annotations or sophisticated preprocessing pipelines, thereby reducing their flexibility, scalability, and generalization performance.

Recently, diffusion models have shown remarkable performance in image generation. Owing to their superior generation quality and stable optimization behavior, they have also been widely adopted for font synthesis. Nevertheless, in few-shot scenarios, diffusion-based font generation still faces notable limitations. During the early denoising stages, structural cues can be easily overwhelmed by noise-dominated features, which weakens the preservation of key character structures and leads to structural deviations or stroke deformation in the final outputs. Moreover, in many existing methods, structural constraints are mainly imposed at the reconstruction or output stage, with insufficient interaction with intermediate feature alignment. At the same time, the commonly used single-path content aggregation strategy is often inadequate for simultaneously modeling the global layout and local stroke details required for complex character generation [2]. As shown in Figure 1, the proposed method can generate visually consistent and structurally coherent characters under few-shot settings, demonstrating its ability to effectively capture both style characteristics and content structures.



Figure 1. Characters synthesized by the proposed method.

To overcome the above limitations, we propose a skeleton-aware deformable alignment framework (SDA-Font) for few-shot font generation. Specifically, the proposed framework introduces skeleton-guided supervision into the diffusion generation process to strengthen the modeling of character topology and structural continuity. A structure-constrained deformable content alignment module is further introduced to improve feature correspondence while suppressing unreasonable deformation, thereby yielding more stable and accurate content representations. In addition, a multi-module content aggregation strategy is developed to integrate structural and semantic cues from multiple representation levels, enabling more effective modeling of both global layout and local details. As a result, the proposed method significantly improves structural consistency and detail fidelity, especially for complex character generation under few-shot conditions.

Beyond improving generation quality, future research should also pay attention to the security and ethical implications of font generation technologies. As generated glyphs become increasingly realistic, issues related to font copyright protection, multimedia content authenticity verification, and synthetic forgery detection deserve further investigation.

Integrating font generation models with security-oriented mechanisms may help promote the responsible deployment of such technologies in practical applications.

Our major contributions can be summarized as follows:

- We propose a skeleton-aware structural supervision scheme by introducing a skeleton prediction branch and a skeleton-aware loss, which effectively regularizes character topology and stroke continuity during generation.
- We develop a structure-constrained deformable content alignment mechanism that preserves alignment flexibility while suppressing unreasonable deformation, thereby improving the structural stability of complex character generation.
- We design a multi-module content aggregation strategy to jointly exploit global structural information and local fine-grained features, thereby enhancing font generation under few-shot settings.

The remainder of this paper is organized as follows. Section 2 reviews the related work on few-shot font generation and diffusion-based font synthesis. Section 3 presents the proposed SDA-Font framework and its key components. Section 4 describes the experimental settings and provides comprehensive quantitative and qualitative evaluations. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

2.1. Image-to-Image Translation

Font generation is commonly formulated as an image-to-image translation (I2I) problem, where the goal is to learn the mapping from source glyphs to target glyphs. Early approaches were predominantly based on fully paired supervised frameworks, such as Pix2Pix [3] and zi2zi [4], which employ conditional generative adversarial networks for font style transfer. Although these methods can achieve satisfactory generation quality given sufficient paired training data, they rely heavily on large-scale aligned datasets and thus suffer from limited scalability and generalization.

To reduce the dependence on paired data, later studies introduced unsupervised I2I methods, which enable cross-domain translation via cycle-consistency learning. However, when adapted to font generation, these methods still encounter problems such as unstable structural alignment and loss of fine-grained details. Some studies further explored latent-space constraints. For instance, Few-Shot Unsupervised Image-to-Image Translation (FUNIT) [5] learns a shared latent space to model the joint distribution of different domains. Nevertheless, such strategies remain insufficient for preserving local structures in complex characters.

These limitations largely stem from the intrinsic discrepancy between font style transfer and general image translation. In generic I2I tasks, style is typically represented by global attributes such as color and texture, whereas font style depends much more strongly on local stroke morphology, spatial composition, and structural coherence. As a result, even slight changes in stroke patterns may noticeably degrade the authenticity and readability of generated characters. Consequently, directly adopting generic I2I models is often inadequate for capturing the structural characteristics of complex glyphs.

Motivated by this observation, recent font generation studies have increasingly moved toward the content–style disentanglement paradigm, forming two representative research directions in few-shot font generation. One direction focuses on global style representation, as exemplified by AGIS-Net [6] and DG-Font [7], but often shows limited capability in preserving local details. The other direction emphasizes local style modeling, as represented by LF-Font [8] and CG-GAN [9], which can better maintain detail fidelity but may weaken

global style consistency. Overall, this evolution reflects a broader shift in font generation research from generic image translation toward structure-aware representation learning.

2.2. Few-Shot Font Generation

Few-shot font generation aims to synthesize a complete target font library from only a small number of reference glyphs. It is typically formulated as an image-to-image (I2I) translation problem, whose key challenge lies in disentangling and transferring content structure and style characteristics from limited reference samples. Because this task requires both structural consistency and style transferability to be maintained simultaneously, it is intrinsically constrained by strong structural priors.

Early studies mainly followed the content–style disentanglement paradigm. EMD [10] and AGIS-Net, for instance, realize cross-font transfer by separately modeling content and style representations, whereas DG-Font introduces deformable convolution to strengthen the modeling of complex stroke morphologies. Deep-CalliFont [11] further improves the stability of complex Chinese character generation through a dual-network architecture. Nevertheless, these methods usually depend on a large amount of paired training data and therefore show limited generalization performance in few-shot scenarios or when handling structurally complex characters.

To reduce reliance on paired samples, some studies have explored few-shot font generation methods based on global feature representation. FontGAN [12] and VAE-Font achieve style transfer and glyph synthesis through a shared latent space. However, their global aggregation strategies tend to weaken the representation of local strokes and fine-grained structural details, which limits their ability to model complex characters effectively.

More recently, component-level and stroke-level modeling methods have attracted increasing attention. ZiGAN [13], LF-Font, and VQ-Font [14] improve structural controllability through component decomposition and recombination. Despite their effectiveness in enhancing structural expressiveness, these methods often rely on additional component annotations or sophisticated preprocessing procedures, which restrict their scalability and practical applicability. In summary, existing few-shot font generation methods still exhibit notable limitations in structural preservation and style transfer. This, in turn, provides strong motivation for developing more fine-grained structure-aware modeling strategies and conditional generation frameworks.

2.3. Diffusion Models

Diffusion models were first introduced by Sohl-Dickstein2015 [15], who formulated data generation as a multi-step Markovian forward diffusion process. In this framework, Gaussian noise is progressively added to real data, while a corresponding reverse denoising process is learned to recover the underlying data distribution. Building upon this formulation, Ho et al. [16] proposed the Denoising Diffusion Probabilistic Model (DDPM), which reformulates model training as a noise prediction task and significantly improves both training stability and generation quality. Owing to their discriminator-free formulation, diffusion models can effectively avoid issues such as mode collapse and have therefore attracted widespread attention in image generation and image restoration.

As research has progressed, condition-guided diffusion models have further broadened the applicability of this paradigm by enabling generation under specified priors and constraints. Existing studies incorporate conditional information, such as text and images, into the reverse denoising process, thereby achieving high-quality image generation and restoration under multimodal guidance. Nevertheless, these methods mainly rely on global conditions for guidance and still face difficulties in effectively controlling fine-grained structural details in tasks with strong structural constraints.

Due to their excellent generative capability and stable optimization behavior, diffusion models have also recently been introduced into font generation. Diff-Font [17] and FontDiffuser [18] employ conditional diffusion frameworks to achieve few-shot font generation and style transfer, while MSD-Font [19] further enhances generation stability in multi-font scenarios. Despite these advances, existing diffusion-based font generation methods remain vulnerable to noise-dominated features during the early diffusion stages, which may progressively degrade critical structural cues. As a result, their ability to preserve structural integrity is still limited, especially when generating complex characters under few-shot conditions.

Recent advances in multimedia technologies have also raised increasing concerns regarding content authenticity, copyright protection, and information security. Beyond improving generation quality, ensuring the trustworthy and responsible deployment of generative technologies has become an important research direction, particularly in applications involving content creation and digital media [20].

2.4. Attention Mechanisms

Attention mechanisms improve feature representation by adaptively reweighting different features and guiding models to focus on the most informative content. Due to their effectiveness, they have been widely adopted in a variety of vision tasks, such as image classification, object detection, and semantic segmentation. Depending on the dimension along which dependencies are modeled, attention mechanisms can be broadly categorized into channel attention, spatial attention, branch attention, and self-attention. Among these variants, self-attention is particularly effective at capturing long-range dependencies across different spatial positions and therefore shows distinct advantages in modeling complex structures.

Since being introduced into the Transformer architecture by Vaswani et al., self-attention has enabled each element to interact with the entire sequence, thereby significantly improving both representational capacity and computational parallelism. As a result, it has become a fundamental building block in modern vision generation models. Later studies further integrated attention mechanisms with convolutional architectures to jointly exploit local perceptual patterns and global contextual dependencies, yielding encouraging results in generation tasks. For instance, RDGAN [21] uses attention mechanisms to extract coarse-grained stroke structures from source characters and transfer them to target styles. HWT [22] incorporates Transformer modules to model semantic associations between images and text, enabling stylized handwritten English text image generation. DeepVecFont-v2 [23] further enhances the stability of modeling long and complex contours through attention-based design.

For generation tasks with strong structural constraints, attention mechanisms are frequently employed to strengthen the modeling of critical regions and structural relationships. Nevertheless, existing methods mainly emphasize global dependency modeling and remain insufficient in terms of multi-scale feature collaboration and fine-grained structural preservation. These limitations provide motivation for subsequent exploration of structure-aware and condition-guided attention mechanisms.

3. Methodology

The overall architecture of the proposed SDA-Font framework is shown in Figure 2. Given a small number of reference glyphs from the target font, the model is designed to generate unseen characters while preserving both structural consistency and stylistic coherence. Specifically, the reference glyphs are used to extract style-related representations, whereas the content image is utilized to capture hierarchical structural features. To further

improve structural preservation and strengthen the interaction between style and content, the proposed framework introduces three key components: a skeleton-aware supervision branch, a structure-constrained deformable content alignment module (SCDCA), and a multi-module content aggregation strategy. These components are jointly embedded into the denoising generation network, thereby enhancing structural consistency, style fidelity, and the overall visual quality of the synthesized characters. More specifically, the skeleton-aware branch provides explicit structural supervision to encourage the learning of stable glyph topological representations. The structure-constrained deformable content alignment module performs flexible matching between reference features and target features by predicting spatial offsets, thereby improving the alignment of structurally relevant content. Meanwhile, the multi-module content aggregation strategy integrates structural and semantic cues from multiple representation levels, which further enhances structural consistency and fine-grained detail modeling in the generated characters.

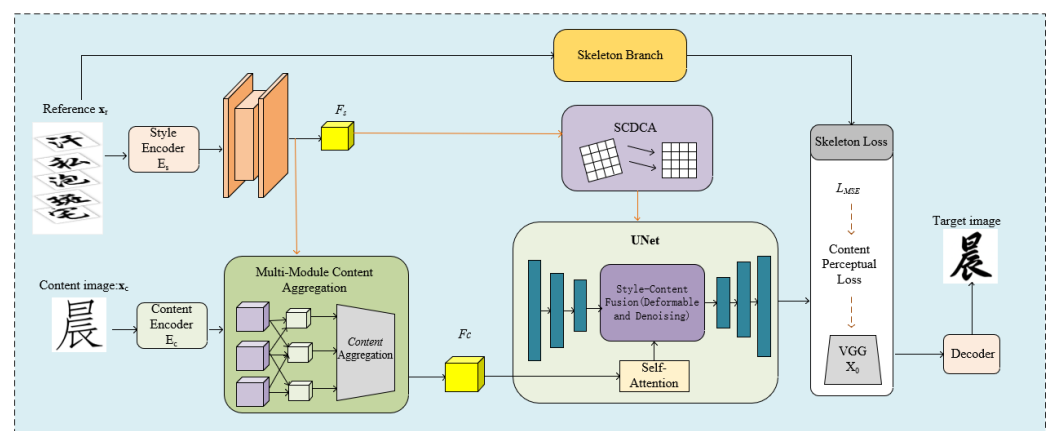


Figure 2. Overview of the proposed SDA-Font framework. The reference glyph image X_r provides style guidance, whereas the content image X_c supplies structural information. The skeleton-aware supervision scheme, the SCDCA module, and the multi-module content aggregation strategy are jointly integrated into the U-Net-based generation network to facilitate few-shot font generation.

3.1. Skeleton-Aware Loss

In few-shot font generation, pixel-level reconstruction losses and diffusion denoising prediction losses mainly constrain local appearance consistency, but are insufficient to explicitly preserve the stroke topology and geometric continuity of characters. This limitation becomes particularly pronounced when generating complex Chinese characters or under extremely limited reference conditions, where the model is prone to artifacts such as stroke disconnection, stroke adhesion, and structural distortion. To alleviate these issues, we introduce a skeleton-aware branch and further design a corresponding skeleton-aware loss to explicitly regularize structural consistency at the skeleton level.

3.1.1. Skeleton Representation and Skeleton Prediction Branch

Given a target glyph image $x \in [0, 1]^{H \times W}$, its centerline skeleton is first extracted through binarization followed by morphological skeletonization. Formally, we introduce a skeleton extraction operator $S(\cdot)$, which sequentially performs normalization, threshold-based segmentation, and skeleton thinning to obtain the skeleton representation:

$$s = S(x), \quad s \in [0, 1]^{H \times W}. \quad (1)$$

Specifically, the glyph image is first normalized to the range $[0, 1]$ and binarized using a fixed threshold of 0.5. The resulting binary mask is then processed by morphological skeletonization to obtain the centerline representation of character strokes. To reduce the

influence of isolated noisy pixels and extremely thin stroke fragments, Gaussian smoothing with $\sigma = 0.8$ is further applied to the extracted skeleton map. In our experiments, the skeleton extraction process remained stable within a reasonable threshold range around 0.5 and did not exhibit noticeable sensitivity to threshold variations.

Although the proposed method relies on skeleton maps extracted by morphological skeletonization, the framework is not highly sensitive to minor extraction errors. First, the extracted skeletons are only used as auxiliary structural supervision rather than strict generation targets. Second, Gaussian smoothing is applied to alleviate the influence of isolated noisy pixels and fragmented skeleton structures. Therefore, moderate inaccuracies in skeleton extraction mainly affect local supervision strength but do not substantially alter the overall generation process. Empirically, we observed stable generation performance under different threshold settings, suggesting that the proposed method exhibits reasonable robustness to imperfect skeleton priors.

The resulting skeleton characterizes the central paths of strokes and their connectivity, thereby providing an effective representation of the global structural characteristics of the glyph.

To explicitly impose skeleton-level structural constraints during the diffusion training process, we introduce a lightweight skeleton prediction branch into the U-Net backbone. This branch predicts a skeleton probability map $\hat{s}$ from the intermediate U-Net [24] features, which captures the structural characteristics of the currently generated glyph. In this way, skeleton information is incorporated as an explicit supervisory signal, encouraging the network to learn more stable and structurally consistent representations.

3.1.2. Skeleton-Aware Loss and Joint Optimization

To constrain the structural consistency between the predicted skeleton $\hat{s}$ and the ground-truth skeleton s , we design a skeleton-aware loss and jointly optimize it together with the original diffusion objective. Specifically, the proposed loss is formulated as a weighted combination of the binary cross-entropy loss and the Dice loss:

$$\mathcal{L}_{skel} = \lambda_{bce} \mathcal{L}_{BCE}(\hat{s}, s) + \lambda_{dice} \mathcal{L}_{Dice}(\hat{s}, s) \quad (2)$$

The Dice loss is defined as

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i \hat{s}_i s_i + \epsilon}{\sum_i \hat{s}_i + \sum_i s_i + \epsilon} \quad (3)$$

where ϵ denotes a smoothing constant introduced to alleviate the class imbalance caused by sparse skeleton pixels.

During training, the skeleton-aware loss serves as an auxiliary structural regularization term and is jointly optimized with the denoising objective. Accordingly, the overall training loss is given by

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_{skel} \mathcal{L}_{skel} \quad (4)$$

Here, $\mathcal{L}_{diff}$ represents the standard denoising objective of the diffusion model. By incorporating the skeleton-aware loss, the proposed model is able to maintain style transfer capability while significantly enhancing the structural stability and stroke connectivity of the generated glyphs.

3.2. Structure-Constrained Deformable Content Alignment

In font generation, content alignment must not only facilitate style transfer but also preserve the topological consistency and structural readability of characters. Conventional

alignment strategies based on fixed convolutions or global attention [25] lack explicit modeling capability for local geometric deformation. Consequently, they often suffer from structural distortions, such as stroke disconnection and stroke misalignment, particularly when dealing with complex glyphs or few-shot scenarios.

To overcome these limitations, we propose a structure-constrained deformable content alignment mechanism (SCDCA), as shown in Figure 3. The proposed module enhances spatial feature alignment under structural constraints. Specifically, the fused feature F_r is first normalized, after which an offset-and-weight predictor is used to adaptively estimate deformable sampling [26] parameters. Based on these parameters, content features are dynamically aligned to obtain the aligned feature F_a . Through this deformable alignment process, the model can more effectively match local stroke regions, thereby improving the alignment precision of content representations.

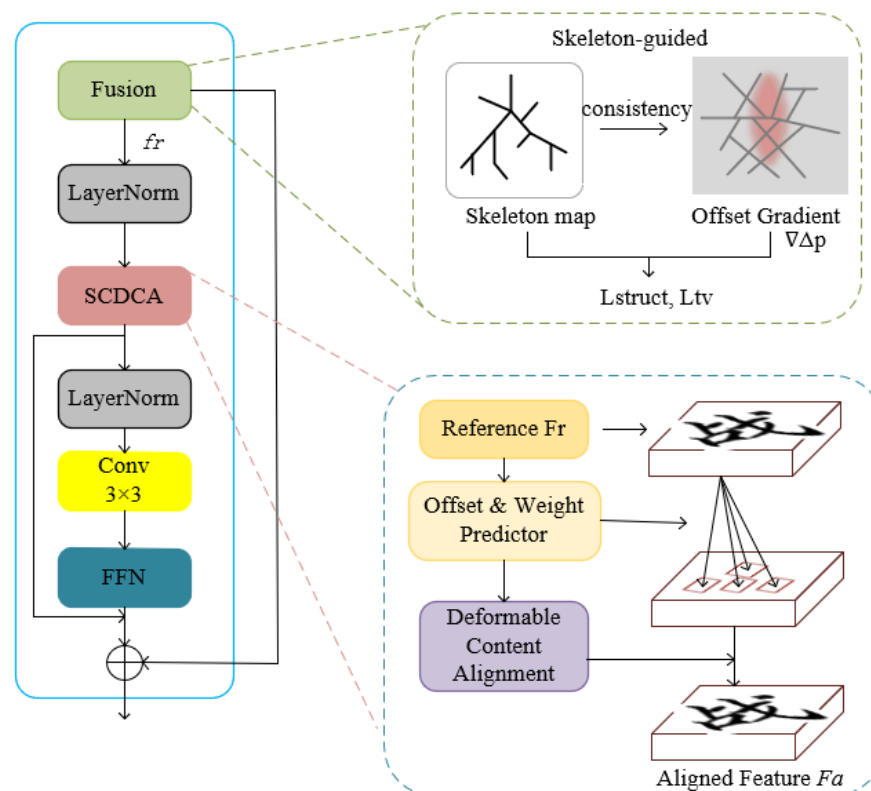


Figure 3. Structural illustration of the proposed SCDCA module.

To prevent excessive deformation during feature alignment, we further introduce a skeleton-guided branch, in which the skeleton map is used as a structural prior to regularize the predicted offsets. Specifically, a structural consistency loss L_{struct} and a total variation loss L_{tv} are employed to constrain offset variations so that they better conform to the topological structure and spatial distribution of character strokes, thereby suppressing structural damage caused by arbitrary displacement. After the alignment stage, the features are further refined by a 3×3 convolution and a feed-forward network, and are then combined with the input features via a residual connection. Overall, the proposed SCDCA module improves local content alignment while contributing to the structural stability of the generated glyphs.

3.2.1. Deformable Content Alignment

Let the reference glyph feature map and the target glyph feature map be denoted by

$$F_r \in \mathbb{R}^{C \times H \times W}, \quad (5)$$

and

$$F_t \in \mathbb{R}^{C \times H \times W}, \quad (6)$$

respectively.

Conventional similarity-based feature fusion methods typically assume strict spatial correspondence between the reference features and the target features, and therefore exhibit limited adaptability to local geometric variations. To overcome this limitation, we introduce a deformable content alignment mechanism that explicitly models local deformation. Specifically, a spatial offset field $\Delta p \in \mathbb{R}^{H \times W \times 2}$ and a corresponding aggregation weight map $A \in \mathbb{R}^{H \times W}$ are predicted to adaptively adjust the sampling positions of reference features in the spatial domain, thereby enabling fine-grained content alignment. The aligned feature is formulated as

$$F_a(\mathbf{p}) = \sum_{k=1}^K A_k(\mathbf{p}) \cdot F_r(\mathbf{p} + \Delta p_k(\mathbf{p})), \quad (7)$$

where K denotes the number of sampling points, and $\mathbf{p}$ is the spatial position index. In this way, the proposed mechanism allows the network to adaptively adjust the geometric layout of reference features according to the local structural characteristics of the target glyph, thereby achieving more accurate alignment of stroke trajectories and spatial relationships.

3.2.2. Structural Consistency Constraint

Although deformable alignment improves the flexibility of local feature matching, excessively unconstrained deformation may damage the topological structure of characters, resulting in artifacts such as stroke disconnection and structural distortion. To suppress unreasonable spatial deformation while preserving structural consistency, we introduce a structural consistency constraint into the offset learning process.

Specifically, let the structure guidance map extracted by the skeleton-aware branch be denoted by

$$S \in \mathbb{R}^{H \times W} \quad (8)$$

Based on this guidance, we impose a structural regularization term on the spatial gradients of the offset field, which is formulated as

$$\mathcal{L}_{struct} = \sum_p \|\nabla \Delta p(p)\|_2 \cdot S(p) \quad (9)$$

where this term constrains the deformation gradients within skeleton regions, thereby preserving the continuity and topological consistency of character strokes. Furthermore, to encourage the overall smoothness of the deformation field, we introduce an additional total variation regularization term:

$$\mathcal{L}_{tv} = \sum_p (\|\nabla_x \Delta p(p)\|_2 + \|\nabla_y \Delta p(p)\|_2) \quad (10)$$

Together, these constraints effectively suppress invalid deformation in non-structural regions without significantly compromising the model's expressive flexibility, thereby improving the stability of the alignment process.

3.2.3. Integration and Training Strategy of the Alignment Module

The proposed SCDCA module is integrated into the feature extraction network of the diffusion model, where content features are progressively aligned across multiple scales. During training, the alignment module is jointly optimized together with the skeleton-aware branch. The overall alignment objective is formulated as

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{recon}} + \lambda_s \mathcal{L}_{\text{struct}} + \lambda_t \mathcal{L}_{\text{tv}} \quad (11)$$

Here, $\mathcal{L}_{\text{recon}}$ denotes the feature reconstruction loss, which ensures that the aligned content features remain semantically consistent with the target glyph. The coefficients λ_s and λ_t are balancing factors used to control the trade-off between structural regularization and deformation smoothness.

Through joint optimization of these objectives, the proposed framework achieves flexible yet precise local content alignment while maintaining the overall structural plausibility of the generated characters.

3.3. Multi-Module Content Aggregation

In few-shot font generation, simultaneously preserving global structural consistency and local stroke-level details remains a challenging problem. Existing methods usually fuse content and style features through simple feature concatenation or a single attention mechanism [27]. Although these approaches can capture partial appearance information, they often fail to fully exploit the complementary cues among different structural representations [28], which may result in structural distortion or detail degradation when generating complex Chinese characters [29]. To overcome this limitation, we propose a multi-module content aggregation strategy (MMCA) to integrate complementary content information extracted from different modules into a unified representation, thereby constructing more complete, stable, and discriminative content features.

Specifically, the proposed strategy models character content from two complementary perspectives. The first is a global structural feature $F_g \in \mathbb{R}^{C \times H \times W}$, extracted from the multi-scale feature maps of the content encoder, which mainly describes the overall topological layout, stroke distribution, and spatial organization of the character. The second is a local detail feature $F_l \in \mathbb{R}^{C \times H \times W}$, jointly provided by the skeleton-aware branch and the local structure modeling module, which captures fine-grained cues such as stroke endpoints, turning points, connection regions, and boundary contours. In this way, the former preserves global structural stability, while the latter enhances sensitivity to local structural variations.

Since these two feature maps share the same spatial resolution, they are first concatenated along the channel dimension [30] to obtain a joint content representation:

$$F_{\text{cat}} = [F_g; F_l] \quad (12)$$

where $[\cdot; \cdot]$ denotes channel-wise concatenation.

To further enhance cross-level content interaction and suppress redundant feature responses, we employ a feature mapping function $\phi(\cdot)$ to compress and fuse the concatenated feature, yielding the final aggregated representation:

$$F_{\text{mca}} = \phi(F_{\text{cat}}) \quad (13)$$

Here, $\phi(\cdot)$ denotes a feature mapping operation composed of convolutional transformations and nonlinear activations.

Through this design, MMCA effectively integrates global topological information and local fine-grained details within a unified representation space, allowing the model to more accurately recover subtle stroke variations while maintaining overall structural stability. Compared with single-path content modeling strategies, the proposed module improves the completeness and robustness of content encoding, thereby providing more reliable content support for subsequent font generation. Its effectiveness in enhancing structural preservation and generation quality is further validated by the ablation experiments.

4. Experiments

4.1. Font Dataset

To perform a systematic evaluation, we constructed a large-scale font dataset covering diverse writing styles. The training set contains 400 seen fonts, each associated with 900 commonly used Chinese characters. For testing, we designed two benchmark subsets: (1) a Seen Fonts test set containing 400 randomly selected visible characters, among which 300 appeared during training; and (2) an Unseen Fonts test set consisting of 40 font styles excluded from training and their corresponding character sets, which is used to evaluate cross-font generalization.

In addition, to further assess robustness under more challenging conditions, we built a harder test set containing 100 font styles and 200 characters different from those in the main dataset, all of which are unseen during training. These unseen characters are generally more structurally complex and stroke-rich, thus providing a stricter evaluation protocol. All characters were selected according to the Stroke Order Standard for Commonly Used Chinese Characters (2021) issued by the National Language Commission of the Ministry of Education of China, ensuring broad coverage of the major structural categories of commonly used Chinese characters. For fair comparison, all character images were resized to 96×96 pixels.

4.2. Implementation Details

Different optimization strategies were adopted for different modules. For style-related modules, we used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$; for content-related modules, RMSprop was employed with $\alpha = 0.999$. The batch size was set to 16, and the model was trained for 440,000 iterations on a single NVIDIA RTX 4090 GPU, taking approximately 30 h. The initial learning rate was set to 1×10^{-4} with a linear decay schedule, and the weight decay was fixed at 1×10^{-2} . To enhance controllability and stability during generation, classifier-free guidance was introduced in the sampling stage. During training, the content image x_c and style image x_s were replaced with constant values (e.g., 255) with a certain probability to simulate unconditional inputs. The sampling process is defined as

$$\epsilon_{\theta}(x_t, t, x_c, x_s) = (1 - s) \epsilon_{\theta}(x_t, t, 0, 0) + s \epsilon_{\theta}(x_t, t, x_c, x_s) \quad (14)$$

where s is the guidance scale, set to 7.5 in all experiments. For efficient inference, DPM-Solver++ [31] was adopted with 20 sampling steps, providing a good balance between generation fidelity and computational overhead.

The diffusion model was trained using 1000 forward diffusion timesteps following the standard DDPM setting. All glyph images were resized to 96×96 pixels before training and evaluation. For skeleton extraction, glyph images were first binarized using a fixed threshold of 0.5 and subsequently processed by morphological skeletonization to obtain centerline representations. To improve training stability, Gaussian smoothing with $\sigma = 0.8$ was further applied to the extracted skeleton maps. The weighting coefficients of the skeleton-aware loss were empirically set to $\lambda_{\text{skel}} = 0.3$, $\lambda_{\text{bce}} = 1.0$, and $\lambda_{\text{dice}} = 1.0$. For the

deformable alignment module, the structural regularization coefficient was set to 0.1. The spatial offsets and aggregation weights were predicted by an offset-and-weight predictor and jointly optimized with the diffusion network during training.

For the proposed SCDCA module in Equation (11), the structure alignment loss and smoothness regularization loss were weighted by $\lambda_s = 1.0$ and $\lambda_t = 0.1$, respectively. In addition, the offset magnitude regularization term was assigned a weight of 0.01. These hyperparameters were selected according to preliminary validation experiments and remained fixed throughout all experiments.

Different optimizers were adopted according to the characteristics of different feature branches. The style-related modules mainly learn global appearance statistics and style distributions, for which Adam provides stable adaptive moment estimation and faster convergence. In contrast, the content-related modules focus on preserving structural consistency and fine-grained stroke topology. RMSprop was employed because it provides smoother updates under sparse and highly structured gradients, which helps reduce local oscillations during content alignment and skeleton-aware supervision. Empirically, we observed stable convergence throughout training and consistent generation quality under this optimization configuration. Therefore, this setting was retained in all reported experiments.

To evaluate the computational efficiency of the proposed method, we report the parameter count and inference time per character, and compare our method with representative baseline models. All methods were evaluated under identical experimental settings to ensure a fair comparison. Inference time is averaged over multiple generated characters after a warm-up phase on the same GPU. The results are summarized in Table 3.

4.3. Comparison with State-of-the-Art Methods

To comprehensively validate the effectiveness of the proposed method, we compare it with several representative font generation approaches, including FS-Font [32], CF-Font [33], Diff-Font, FontDiffuser, MSD-Font, and GC-Font [34]. These methods improve font generation performance from different perspectives, such as local pattern modeling, shared latent space alignment, stroke-component-conditioned generation, multi-scale content aggregation, multi-stage diffusion generation, and global contextual constraints, thus providing strong representativeness and comparability.

Specifically, FS-Font enhances style consistency by exploring fine-grained local patterns in reference glyphs as well as the spatial correspondence between reference and content glyphs. CF-Font maps the style and content features of source and target fonts into a shared latent space and employs distribution distance as a reconstruction constraint, thereby improving the accuracy of font style transfer. Diff-Font introduces stroke-component information as auxiliary conditions into a diffusion-based generation framework to strengthen fine-grained structural modeling. FontDiffuser improves the reconstruction quality of complex characters and scenarios with large style discrepancies by incorporating a multi-scale content aggregation module and a style contrastive refinement mechanism. MSD-Font adopts a multi-stage dual-branch diffusion architecture to progressively model structural, stylistic, and detailed information under limited reference samples, thereby enhancing few-shot font generation performance. GC-Font further introduces global contextual feature modeling and integrates skeleton consistency and edge consistency constraints, which effectively improve the structural integrity and overall generation quality of complex characters.

4.3.1. Fair Comparison Protocol

To ensure a fair and unbiased comparison, all competing methods were evaluated under the same experimental protocol. Specifically, all methods were trained and tested using the identical dataset splits described in Section 4.1. For few-shot font generation,

the same number of reference glyphs was provided to each method during inference. In addition, all glyph images were resized to a uniform resolution of 96×96 pixels, and the quantitative results were computed using the same evaluation metrics, including L1, SSIM, FID, and LPIPS. Whenever publicly available, official implementations and recommended hyperparameter settings were adopted for all baseline methods. For methods without publicly released code, we reproduced them as faithfully as possible according to the descriptions provided in the original papers.

Furthermore, all baseline methods were evaluated under our experimental setup rather than using the results reported in the original publications. Whenever official implementations were available, the corresponding models were retrained on our dataset using the same training and testing splits described in Section 4.1. This protocol ensures a fair, reproducible, and unbiased comparison across all competing methods.

4.3.2. Quantitative Comparison

To comprehensively assess the effectiveness of the proposed method for few-shot font generation, we performed quantitative comparisons with competing approaches under both SFUC and UFUC settings using four widely adopted metrics, namely L1, SSIM [35], FID [36], and LPIPS [37]. The corresponding results are reported in Table 1. Overall, the proposed method achieves the best performance in both settings, indicating its superior capability in character structure preservation, image quality enhancement, and perceptual consistency.

Table 1. Quantitative Comparison on the SFUC and UFUC Datasets.

Methods	SFUC				UFUC			
	L1 ↓	SSIM ↑	FID ↓	LPIPS ↓	L1 ↓	SSIM ↑	FID ↓	LPIPS ↓
FS-Font	0.1939	0.4137	29.3078	0.4789	0.2177	0.4256	21.5103	0.3205
CF-Font	0.1892	0.4244	28.7133	0.4178	0.2152	0.4366	20.5706	0.2851
Diff-Font	0.1765	0.4366	27.8421	0.3658	0.2106	0.4421	18.4689	0.2752
MSD-Font	0.1644	0.4597	24.8534	0.3585	0.1993	0.4534	17.5924	0.2427
FontDiffuser	0.1382	0.4605	23.2893	0.2903	0.1805	0.4667	16.1658	0.2275
GC-Font	0.1290	0.4638	21.4350	0.2782	0.1627	0.4751	15.5443	0.2124
Our	0.1110	0.4705	20.4940	0.2140	0.1592	0.4950	14.3204	0.1929

4.3.3. Qualitative Comparison

The qualitative results are presented in Figure 4. We selected 20 Chinese characters for visual comparison under both SFUC and UFUC settings, including FS-Font, CF-Font, Diff-Font, MSD-Font, FontDiffuser, GC-Font, and the proposed method.

As can be observed, FS-Font and CF-Font tend to produce unstable structures and insufficient stroke details when generating complex characters. Although Diff-Font can recover the overall contour of characters, it still suffers from blurring and missing fine details in structurally complex regions. MSD-Font performs relatively well on simple characters, but its local stroke restoration remains inadequate for complex ones. FontDiffuser and GC-Font further improve structure preservation and style representation; however, noticeable detail deviations still appear in unseen-character and complex-structure scenarios. In contrast, the proposed method achieves superior visual results under both settings, preserving character structural integrity more accurately while better recovering complex stroke details and target font styles. These results demonstrate its stronger robustness and generalization capability in few-shot font generation.

Source	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
FS-Font	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
CF-Font	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
Diff-Font	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
MSD-Font	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
FontDiffuser	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
GC-Font	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
Our	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育
Target	安	错	笔	度	遍	曾	登	根	很	击	解	梦	景	南	势	索	希	新	征	育

Source	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
FS-Font	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
CF-Font	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
Diff-Font	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
MSD-Font	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
FontDiffuser	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
GC-Font	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
Our	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴
Target	冲	拙	霖	枫	邓	荒	言	担	军	蜂	轮	懵	柯	董	龚	舟	估	酒	韩	贴

Figure 4. Qualitative Comparison on SFUC and UFUC.

More detailed observations can be made from several representative examples in Figure 4. For instance, in the SFUC setting, several complex Chinese characters with dense stroke interactions and intricate structural layouts are selected for comparison. Compared with baseline methods, the proposed method preserves stroke continuity more effectively and avoids broken or merged strokes, resulting in clearer character topology. In contrast, FS-Font and CF-Font occasionally exhibit stroke discontinuities and local structural inconsistencies, while Diff-Font tends to generate blurred stroke boundaries in densely connected regions. Similarly, in the UFUC setting, challenging unseen characters require accurate modeling of both global structure and local stroke details. The proposed method maintains better alignment between stroke components and overall character layout, thereby reducing deformation artifacts and improving structural fidelity. Although FontDiffuser and GC-Font generate visually plausible results, slight deviations can still be observed in complex stroke intersections and fine-grained details. Benefiting from the skeleton-aware supervision and structure-constrained deformable alignment mechanism, the proposed method achieves superior preservation of character structure and stroke connectivity, particularly for unseen and structurally complex glyphs.

4.4. Ablation Studies

The ablation experiments were conducted on the UFUC dataset, and the quantitative results are reported in Table 2. To verify the effectiveness of each module, we conduct qualitative ablation experiments by removing Skel, SCDCA, and MMCA individually from the full model. To ensure a fair evaluation, each component was independently removed from the full model while all other modules and training settings were kept unchanged. Specifically, “w/o Skel” denotes the variant without the Skeleton-Aware Loss, “w/o SCDCA” denotes the variant without the Structure-Constrained Deformable Content Alignment module, and “w/o MMCA” denotes the variant without the Multi-Module Content Aggregation strategy. This design allows the individual contribution of each proposed component to be isolated and quantitatively evaluated. Table 2 reports the results

of single-factor ablation experiments, where only one component is removed at a time while all remaining modules and training settings are kept unchanged.

Table 2. Ablation study results obtained by independently removing each proposed component from the full model.

Models	L1 ↓	SSIM ↑	FID ↓	LPIPS ↓
Our	0.1592	0.4950	14.3204	0.1929
w/o Skel	0.1782	0.4196	16.4102	0.2422
w/o SCDCA	0.1815	0.4309	17.1005	0.2553
w/o MMCA	0.1926	0.4543	18.6142	0.2198

As shown in Figure 5, the complete model generates results that are closest to the target glyphs in terms of structural integrity, stroke clarity, and style consistency.

Our	地 耕 顾 脉 战 碎
w/o Skel	地 耕 顾 脉 战 碎
w/o SCDCA	地 耕 顾 脉 战 碎
w/o MMCA	地 耕 顾 脉 战 碎
Target	地 耕 顾 脉 战 碎

Figure 5. Comparison of Character Quality Generated by Ablation Models.

After removing Skel, the generated results exhibit reduced global structural stability, and slight deformation or misalignment can be observed in some complex characters. This indicates that the skeleton constraint helps preserve the overall character structure. After removing SCDCA, the local stroke shapes become less natural, and the correspondence between content and style becomes weaker, suggesting that this module contributes to accurate alignment between structural information and stylistic features. After removing MMCA, the generated characters remain generally recognizable, but their fine-detail restoration and style expressiveness are noticeably degraded, demonstrating that multi-scale aggregation is beneficial for feature fusion and detail preservation.

Overall, all three modules make positive contributions to the generation performance. Specifically, Skel mainly improves global structural consistency, SCDCA enhances alignment between content and style, and MMCA strengthens multi-scale feature fusion and fine-grained stroke reconstruction. The joint effect of these three modules enables the full model to achieve the best visual results.

The quantitative ablation results are reported in Table 2. The full model achieves the best overall performance across all evaluation metrics (L1 = 0.1592, SSIM = 0.4950, FID = 14.3204, LPIPS = 0.1929). After removing the skeleton constraint module, L1 and FID increase to 0.1782 and 16.4102, respectively, while SSIM decreases to 0.4196, verifying the critical role of skeleton information in preserving the global character structure. When

SCDCA is removed, LPIPS deteriorates significantly to 0.2553, indicating the necessity of this module for improving local alignment and cross-domain consistency. After removing MMCA, the FID score worsens to 18.6142, demonstrating that multi-scale feature aggregation plays an indispensable role in detail restoration and semantic feature fusion. These quantitative results are consistent with the visual analysis in Figure 4, further confirming the collaborative effectiveness of the proposed modules in structural preservation, local alignment, and detail enhancement.

Computational Efficiency Analysis

To comprehensively evaluate computational efficiency, we report parameter count, inference time per character, and FLOPs for all compared methods. FLOPs are measured as the total number of floating-point operations required for a single forward pass under the same input resolution and diffusion steps, ensuring a fair comparison across different models.

Table 3. Computational complexity comparison with representative methods.

Method	Params (M)	Inference Time (s/char)	FLOPs (G)
Diff-Font	120.52	2.65	4.13
FontDiffuser	138.35	2.48	3.25
GC-Font	145.23	2.25	2.85
Ours	167.85	2.12	2.52

4.5. Cross-Lingual Font Generation

To further evaluate the proposed model in terms of style transfer and generalization across different writing systems, we conduct cross-lingual [38] font generation experiments. In this setting, large-scale paired supervision from the target language is not introduced during training. Instead, a small number of reference glyphs from one language are used as style conditions, while characters from another language are adopted as content inputs for generation. This setting provides a more rigorous evaluation of whether the model can learn transferable style representations and maintain structural consistency across languages.

As shown in Figure 6, two representative cross-lingual generation results are presented, including Korean [39] and Japanese character generation. It can be observed that the proposed model is able to preserve the stroke thickness, writing texture, and overall visual style of the reference font, while maintaining the structural integrity and readability of the target characters. In the Korean generation task, the model produces characters that remain stylistically consistent and visually clear under different reference styles. In the Japanese generation task, the model likewise captures the stylistic characteristics of the reference font and transfers them to the target characters in a coherent manner. The generated results exhibit good consistency in both stroke rendering and overall layout.

These results indicate that the proposed model is capable of learning transferable style representations and achieving effective style transfer across different languages, thereby demonstrating its potential for cross-lingual font generation.

Source Reference	당도떠를그을	いたどのるを	Source Reference
ニ	당도떠를그을	いたどのるを	關
ρ	당도떠를그을	いたどのるを	保
互	당도떠를그을	いたどのるを	川
覚	당도떠를그을	いたどのるを	古
u	당도떠를그을	いたどのるを	樹

Figure 6. Cross-Lingual Generation.

5. Conclusions

In this paper, we proposed SDA-Font, a skeleton-aware deformable alignment framework for few-shot Chinese font generation, to address the challenges of structural distortion, insufficient stroke-detail recovery, and limited generalization ability. Based on a diffusion framework, SDA-Font integrates three key components: skeleton-aware structural supervision, structure-constrained deformable content alignment, and multi-module content aggregation. These components jointly improve structural consistency, detail preservation, and style expressiveness, especially for complex Chinese characters. Extensive experiments show that the proposed method achieves superior performance over existing methods in both quantitative evaluation and qualitative visualization.

Despite the promising results, several limitations remain. First, the introduction of the skeleton-aware supervision branch and the deformable alignment module increases the computational complexity and training cost of the framework. Second, the effectiveness of the proposed method partially depends on the quality of the extracted skeleton representations, which may become less reliable for highly stylized or noisy glyphs. In addition, slight structural deviations may still occur when handling extremely complex character layouts or highly decorative font styles. Furthermore, in extremely sparse-reference scenarios where only one or two reference glyphs are available, insufficient style information may reduce style consistency and fine-grained detail reconstruction. Addressing these limitations will be part of our future work.

In future work, we will investigate lightweight model designs, stronger explicit structural priors, and broader cross-lingual font generation tasks, and robustness evaluation under different reference counts, multiple data splits, and repeated runs to further improve the applicability and generalization of the proposed framework.

Author Contributions: G.Z.: Conceptualization, Methodology, Writing—review & editing. S.W.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing—original draft, Data curation, Visualization. T.J.: Validation, Data curation. J.Y.: Software, Formal analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 61772179), the Open Fund of Hunan Provincial Key Laboratory of Intelligent Information Processing and Application for Hengyang Normal University (Grant No. IIPA20K01), and the Horizontal Research Project of Hengyang Normal University (Grant No. HXKY2022006).

Data Availability Statement: The data presented in this study are openly available at <https://github.com/wuson123/font> (accessed on 27 May 2026).

Acknowledgments: The authors used AI-assisted tools for language polishing, while all scientific content and conclusions were generated and verified by the authors.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Chen, W.; Zhu, G.; Li, Y.; Ji, Y.; Liu, C. Advancements in Chinese font generation since deep learning era: A survey. *arXiv* **2025**, arXiv:2508.06900. [CrossRef]
- Xiao, H.; Fang, Y.; Fu, J.; Guan, L.; Li, M.; Zhou, W.; Zhu, J. Few-shot font generation by frequency feature separation and multi-scale feature aggregation. *Digit. Signal Process.* **2026**, *170*, 105777. [CrossRef]
- Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
- Li, Y.; Duan, J.; Su, X.; Zhang, L.; Yu, H.; Liu, X. Calligraphy generation algorithm based on improved generative adversarial network. *J. Zhejiang Univ. (Eng. Sci.)* **2023**, *57*, 1326–1334.
- Liu, M.-Y.; Huang, X.; Mallia, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
- Gao, Y.; Guo, Y.; Lian, Z.; Tang, Y.; Xiao, J. Artistic glyph image synthesis via one-stage few-shot learning. *ACM Trans. Graph.* **2019**, *38*, 1–12.
- Xie, Y.; Chen, X.; Sun, L.; Lu, Y. DG-Font: Deformable generative networks for unsupervised font generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5130–5140.
- Park, S.; Chun, S.; Cha, J.; Lee, B.; Shim, H. Few-shot font generation with localized style representations and factorization. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2393–2402. [CrossRef]
- 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13482–13491.
- Zhang, Y.; Zhang, Y.; Cai, W. Separating style and content for generalized style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8447–8455.
- Liu, Y.; Lian, Z. DeepCalliFont: Few-shot Chinese calligraphy font synthesis by integrating dual-modality generative models. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 3774–3782. [CrossRef]
- Liu, X.; Meng, G.; Xiang, S.; Pan, C. FontGAN: A unified generative framework for Chinese character stylization and de-stylization. *arXiv* **2019**, arXiv:1910.12604. [CrossRef]
- Wen, Q.; Li, S.; Han, B.; Yuan, Y. Zigan: Fine-grained Chinese calligraphy font generation via a few-shot style transfer approach. In Proceedings of the 29th ACM International Conference on Multimedia; ACM: New York, NY, USA, 2021; pp. 621–629.
- 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 19506–19516.
- Sohl-Dickstein, J.; Weiss, E.A.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2015; pp. 2256–2265.
- Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the Advances in Neural Information Processing Systems 33, Online, 6–12 December 2020; pp. 6840–6851.
- He, H.; Chen, X.; Wang, C.; Liu, J.; Du, B.; Tao, D.; Yu, Q. Diff-Font: Diffusion model for robust one-shot font generation. *Int. J. Comput. Vis.* **2024**, *132*, 5372–5386. [CrossRef]
- 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. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 6603–6611.
- Fu, B.; Yu, F.; Liu, A.; Wang, Z.; Wen, J.; He, J.; Qiao, Y. Generate like experts: Multi-stage font generation by incorporating font transfer process into diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 6892–6901.
- Lin, Y.; Wei, Y.; Chen, D.; Li, Y.; Erkan, U.; Toktas, A.; Gao, S.; Zhang, Y. Cryptanalysis and improvement of a video cryptosystem via chaos and S-box. *ACM Trans. Multimed. Comput. Commun. Appl.* **2026**, *22*, 1–28. [CrossRef]
- Huang, Y.; He, M.; Jin, L.; Wang, Y. RD-GAN: Few/zero-shot Chinese character style transfer via radical decomposition and rendering. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 156–172.
- Bhunia, A.K.; Khan, S.; Cholakkal, H.; Anwer, R.M.; Khan, F.S.; Shah, M. Handwriting transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1086–1094.
- Wang, Y.; Wang, Y.; Yu, L.; Zhu, Y.; Lian, Z. Deepvecfont-v2: Exploiting transformers to synthesize vector fonts with higher quality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18320–18328.

24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
25. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
26. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 15–20 June 2019; pp. 9308–9316.
27. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Style transfer with adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 15–20 June 2019; pp. 6243–6252.
28. Wang, Y.; Zhang, Y.; Wang, X.; Song, Z. MX-Font: Multiple stroke extraction for few-shot font generation. In *Proceedings of the 17th European Conference on Computer Vision*, Tel Aviv, Israel, 23–27 October 2022; pp. 123–139.
29. Li, J.; Liu, Y.; Zhang, T. Few-shot font generation with weak stroke supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 1234–1242.
30. Li, Z.; Chen, S.; Liang, D. SGD-Font: Style and glyph decoupling for one-shot font generation. *Knowl.-Based Syst.* **2025**, *330*, 114600.
31. Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; Zhu, J. Dpm-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv* **2022**, arXiv:2211.01095.
32. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 18–24 June 2022; pp. 7895–7904.
33. Wang, C.; Zhou, M.; Ge, T.; Jiang, Y.; Bao, H.; Xu, W. Cf-font: Content fusion for few-shot font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 17–24 June 2023; pp. 1858–1867.
34. Chen, W.; Zhu, G.; Li, Y.; Ji, Y.; Liu, C. GC-Font: Few-shot font generation via global contextual feature modelling. In *Proceedings of the 36th British Machine Vision Conference*, Sheffield, UK, 24–27 November 2025.
35. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
36. 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 *Proceedings of the Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 4–9 December 2017.
37. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
38. Li, C.; Taniguchi, Y.; Lu, M.; Konomi, S.; Nagahara, H. Cross-language font style transfer. *Appl. Intell.* **2023**, *53*, 18666–18680. [[CrossRef](#)]
39. Park, J.; Lee, J.; Kim, H. Few-shot font generation with deep metric learning and Korean composition prior. In *Proceedings of the 16th International Conference on Document Analysis and Recognition*, Lausanne, Switzerland, 5–10 September 2021; pp. 678–685.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.