

中图法分类号: 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-17

论文引用格式: Li Xiang, Meng Lei, Wu Lei, Meng Xiangxu. Multi-conditional Diffusion for Controllable Zero-shot Artistic Font Generation[J/OL]. Journal of Image and Graphics, XXXX:1-17. DOI: 10.11834/jig.260483. (李祥, 孟雷, 武蕾, 孟祥旭. 面向可控零样本艺术字生成的多条件扩散方法[J/OL]. 中国图象图形学报, XXXX:1-17. DOI: 10.11834/jig.260483.) [DOI: 10.11834/jig.260483]

面向可控零样本艺术字生成的多条件扩散方法

李祥, 孟雷, 武蕾, 孟祥旭

山东大学软件学院, 山东, 济南 250100

摘要: **目的** 零样本艺术字生成需从独立参考样本中重组字符身份、字形结构和视觉字效。针对现有方法整体建模艺术字风格、异构条件协调不足, 导致未见属性组合下结构扭曲和风格偏移的问题, 提出一种面向可控零样本艺术字生成的多条件扩散方法。 **方法** 提出组合式多条件扩散模型, 以独立的字符、字形和字效参考图作为生成条件; 设计跨域特征聚合模块, 以字符特征为结构锚点, 降低字形与字效信息之间的相互干扰; 引入基于自适应实例归一化的跳跃连接, 传递多尺度结构特征与风格统计信息; 采用组合式对比学习约束匹配与不匹配属性组合。为弥补现有数据集在属性多样性、结构复杂度和组合覆盖方面的不足, 并系统评估模型对未见字符、字形和字效组合的泛化能力, 构建包含 921,600 幅图像、30 种字形和 30 种字效的中英文艺术字数据集 SSAFv2。 **结果** 在 SSAFv2 的未见字符—已见风格 (unseen character-seen style, UCSS)、已见字符—未见风格 (seen character-unseen style, SCUS) 和未见字符—未见风格 (unseen character-unseen style, UCUS) 中文子集上, 提出模型的 L1 损失分别为 0.1754、0.1417 和 0.1759, 弗雷歇初始距离 (Fréchet Inception Distance, FID) 分别为 34.58、30.19 和 52.73, 结构相似性 (structural similarity index measure, SSIM) 分别为 0.6351、0.8600 和 0.4870, 整体优于对比方法; 在 Fonts、TE141K-EN 和 SSAFv2-EN 数据集上的表现也优于对比方法。定性结果与用户评价表明, 该方法能够较好地保持复杂字符结构和局部字效细节。 **结论** 将字符、字形和字效作为独立属性进行建模, 并通过跨域特征聚合与组合式对比学习实现多条件协同, 可有效缓解异构属性融合中的特征干扰与错误耦合, 增强模型对属性差异及其组合关系的感知能力, 从而提高未见属性组合下艺术字的结构保真性与风格一致性。本文代码及数据集已在 ScienceDB 公开, 链接为: <https://doi.org/10.57760/sciencedb.013hc>。

关键词: 体生成; 扩散模型; 艺术字数据集; 零样本生成; 组合式对比学习

Multi-conditional Diffusion for Controllable Zero-shot Artistic Font Generation

Li Xiang, Meng Lei, Wu Lei, Meng Xiangxu

School of Software, Shandong University, Jinan 250100, China

Abstract: Objective Zero-shot artistic font generation synthesizes target artistic font images from independent reference samples and supports graphic design and multimedia content creation. The task requires a model to identify, disentangle, and recombine three heterogeneous attributes: character identity, glyph structure, and visual text effect. Character identity determines semantic correctness and topology, glyph structure controls strokes and connectivity, and visual text effect defines texture, shading, color, and background. Because these attributes correspond to discrete semantics, topological

收稿日期: 2026-08-04; 修回日期: 2026-10-08

基金项目: 山东省基础软件重点实验室 (项目编号: 11150004040955); 数字媒体技术教育部工程研究中心; 山东省重点研发计划 (2024TSGC0051);

Supported by: Shandong Key Laboratory of Foundational Software; Engineering Research Center of Digital Media Technology, Ministry of Education; Shandong Provincial Key Research and Development Program

© 中国图象图形学报版权所有

structure, and high-frequency visual patterns, they differ substantially in representation and can interfere with one another during unified generation. Existing methods often model artistic font style as a single holistic factor, depend on adversarial training or attribute-wise supervision, or rely on static masks and predefined templates. Although such methods may perform well on familiar pairings observed during training, they tend to produce missing strokes, structural drift, texture leakage, color contamination, and style deviation when characters, glyphs, and effects are recombined in unseen ways. To address these problems, a multi-condition diffusion method is proposed for controllable zero-shot artistic font generation. The objective is to model character, glyph, and effect information as independent but coordinated conditions for unseen attribute combinations. **Method** A compositional multi-condition diffusion framework, is developed with three condition-specific encoders and a conditional U-Net backbone. Independent character, glyph, and visual-effect reference images are encoded into multi-scale latent features and injected into the denoising process as separate guidance signals. This design allows semantic identity, structural design, and appearance style to jointly control the reverse diffusion trajectory without being compressed into a unified style code. The framework contains three major components. First, a cross-domain feature aggregation module uses character features as structural anchors and introduces channel recalibration to preserve global topology after repeated downsampling. It further applies perpendicular-component attention to remove redundant or interfering components from the joint glyph-effect feature space. This operation reduces the propagation of irrelevant textures from glyph references and unwanted structural patterns from effect references. Second, adaptive instance normalization-based skip connections transfer multi-scale structural features and style statistics from the encoder to the decoder. Means and standard deviations extracted from fused glyph and effect features are used to modulate normalized content features at corresponding decoding stages, enabling the model to retain clear stroke boundaries while reconstructing target colors, textures, and backgrounds. Third, compositional contrastive learning is introduced to improve sensitivity to attribute mismatch. Character, glyph, and effect are treated as separately alignable dimensions. For each target sample, matched combinations are used as positive references, while three negative groups are generated by replacing one attribute at a time. An InfoNCE-based objective pulls generated representations toward the correct composition and pushes them away from combinations containing an incorrect character, glyph, or effect. The final training objective combines denoising loss, perceptual loss, and compositional contrastive loss. To address the limited attribute diversity, structural complexity, bilingual coverage, and combination completeness of existing datasets, the SSAFv2 dataset is constructed for systematic evaluation of unseen combinations. It contains 921,600 artistic font images, 1,024 Chinese and English characters, 30 glyph styles, and 30 visual effects, forming 900 complete glyph-effect categories. The glyph library includes standard printing, creative printing, handwritten styles, and calligraphic styles, while the effect library covers fine textures inspired by natural and industrial materials. SSAFv2 is divided into unseen-character/seen-style, seen-character/unseen-style, and unseen-character/unseen-style subsets to evaluate semantic generalization, compositional style generalization, and their joint difficulty.

Result Experiments are conducted on the SSAFv2, TE141K-EN, and Fonts datasets under a unified evaluation protocol. Performance is measured using L1 loss, Fréchet Inception Distance, and structural similarity, supplemented by qualitative comparison, ablation studies, latent-feature analysis, contrastive-score visualization, and a user study. On the unseen-character/seen-style, seen-character/unseen-style, and unseen-character/unseen-style subsets of SSAFv2, ArtFontDiff obtains L1 losses of 0.1754, 0.1417, and 0.1759; Fréchet Inception Distance scores of 34.58, 30.19, and 52.73; and structural similarity scores of 0.6351, 0.8600, and 0.4870, respectively. These results are overall superior to those of the compared zero-shot artistic font synthesis, image-to-image translation, and diffusion-based font generation methods. On Fonts, TE141K-EN, and SSAFv2-EN, the proposed framework achieves L1 losses of 0.1030, 0.0753, and 0.0380; Fréchet Inception Distance scores of 11.37, 24.21, and 62.27; and structural similarity scores of 0.8897, 0.9194, and 0.9101. Qualitative comparisons show that the proposed model produces clearer structural alignment and more accurate local appearance details than the baselines, particularly for complex unseen Chinese characters, creative glyphs, calligraphic forms, and difficult glyph-effect pairings. Ablation experiments further verify the contribution of each component. Removing channel recalibration causes stroke loss and structural deformation; removing perpendicular-component attention allows irrelevant reference patterns to contaminate the output; removing adaptive instance normalization-based skip connections introduces artifacts and weakens texture fidelity; and removing compositional contrastive learning leads to structural

drift and inconsistent rendering. The complete model achieves an L1 loss of 0.1409, a Fréchet Inception Distance of 62.01, and a structural similarity score of 0.6855 in the ablation setting, outperforming all reduced variants. Latent-space analyses show that compositional contrastive learning produces more compact effect clusters, clearer glyph separation, and more consistent organization by character identity. In the user study, the proposed model receives the highest average scores for both Chinese and English samples. The scores range from 4.36 to 4.50 across the four evaluation dimensions for Chinese results and from 4.66 to 4.78 for English results, confirming improvements in character correctness, glyph similarity, effect similarity, and overall visual quality. **Conclusion** The results support the central argument that controllable zero-shot artistic font generation benefits from independently modeling character identity, glyph structure, and visual effect while coordinating them through cross-domain feature aggregation, multi-scale style modulation, and compositional supervision. This design alleviates feature interference and erroneous coupling among heterogeneous attributes, strengthens the model's sensitivity to attribute differences and composition relationships, and improves structural fidelity and style consistency for unseen combinations. The source code and ssafv2 dataset of this paper are publicly available on Science Data Bank at <https://doi.org/10.57760/sciencedb.013hc>.

Key words: font generation; diffusion models; artistic font dataset; zero-shot generation; compositional contrastive learning

论文引用格式: Li Xiang, Meng Lei, Wu Lei, Meng Xiangxu. Multi-conditional Diffusion for Controllable Zero-shot Artistic

Font Generation [J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260483. (李祥, 孟雷, 武蕾, 孟祥旭. 面向可

控零样本艺术字生成的多条件扩散方法[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260483.)

0 引言

艺术字自动合成在图形设计、视觉传播和多媒体内容创作中具有重要价值。设计者通常需要在保证字符可读性和设计意图的同时,快速生成具有特定字形结构与视觉效果的艺术字图像。本文研究可控零样本艺术字生成任务:给定三张相互独立的参考图像,分别提供目标字符身份、字形结构和艺术字效,模型需要提取并重组三类参考属性,生成同时满足相应语义、结构与外观约束的目标艺术字。其中,字符参考决定生成结果的语义身份及基本拓扑布局;字形参考规定笔画粗细、长宽比例、连通关系、空间排布和收笔形态等结构特征;字效参考则提供颜色、纹理、材质、阴影、渐变及背景等视觉外观信息。这里的零样本设定是指,推理阶段所要求的特定“字符—字形—字效”属性组合未在训练集中出现,模型需要利用已学习的单一属性知识完成新的跨属性组

合,而非复现训练阶段已有的固定配对。由于三类属性分别涉及离散语义、拓扑结构和高频外观细节,其特征分布、空间尺度及作用方式存在明显差异。若在统一生成过程中直接融合这些异构条件,字形和字效信息可能干扰字符骨架,字效中的纹理与颜色也可能错误传播至结构特征,进而造成笔画缺失、结构漂移、纹理污染或风格偏移。因此,该任务的核心挑战是在保持字符可读性与字形完整性的同时,实现三类属性的有效解耦、精确重组以及对未见属性组合的稳健泛化。

现有艺术字生成研究主要沿三条路线展开: Mao等(2023)关注复杂纹理与结构下的字效迁移, Gao等(2019)探索整体风格迁移, Li等(2022, 2023)和 Zhu等(2022)进一步将字形与字效因素解耦。前两类通常将风格作为整体因素建模,难以对结构与渲染进行细粒度控制;解耦式方法更接近零样本组合,但多数仍依赖生成对抗网络(generative adversarial network, GAN)监督,并常按单一属性分别判别。因此,模型在训练集内常见配对上表现较好,却容易在未见组合上出现笔画缺失、结构漂移或颜色污染。生成式零样本研究分别从特征精炼(Chen S等, 2025b)、类别归纳(Chen S等, 2025a)、进化架构搜索(Chen S等, 2024)、扩散增强(Ye等, 2024)和层次语义—视觉对齐(Chen S等, 2021)改进迁移;姿态与服饰生成中的多条件扩散研究也表明,稳健的条件机制是融合异构信号的关键(Shen等, 2023; Shen和Tang, 2024; Shen等, 2025)。然而,这些工作仍难

以直接处理艺术字中离散字符语义、字形拓扑结构和高频字效细节协同的问题。

近期, Diff-Font(He 等, 2024)、FontDiffuser(Yang 等, 2024)、FontStudio(Mu 等, 2024)和 ArtTypo(Liu 等, 2025)利用扩散先验提升了字形与视觉生成的稳定性。但是, FontDiffuser(Yang 等, 2024)、Anything to Glyph(Wang 等, 2023)和 FontAnimate(Fu 等, 2025)等方案仍依赖静态二值蒙版控制轮廓, 或未显式建模字符、字形与字效在未见组合下的重组方式。该问题在中英文双语艺术字中尤为关键, 因为结构保真直接影响可读性。面向这一任务, 构建联合多条件注入和组合学习的框架, 以实现可控零样本生成。

为此, 提出面向可控零样本艺术字生成的多条件组合扩散框架(multi-conditional artistic font diffusion model, ArtFontDiff)。模型以独立的字符、字形和字效参考图作为条件; 在网络层面, 设计跨域特征聚合模块(cross-domain feature aggregation, CFA), 降低异构结构与外观特征在去噪过程中的相互干扰; 在训练层面, 采用组合式对比学习(compositional contrastive learning, CCL), 使生成样本在潜空间中接近匹配属性组合并远离不匹配组合。为在更丰富且贴近应用的设定中评测, 进一步构建 SSAFv2 数据集, 包含 30 种字形、30 种字效、中英文字符覆盖以及完整的属性组合关系。

在 SSAFv2、TE141K(Yang 等, 2021)和 Fonts(Ge 等, 2020)数据集上的实验表明, ArtFontDiff 相比对比方法获得更好的定量与定性结果, 尤其在未见字符、字形和字效组合上更稳定。实验部分还给出消融实验和特征分析, 用于分析 CFA 与 CCL 对组合属性感知能力的影响。主要贡献包括: 1) 提出使用独立字符、字形和字效参考图的多条件扩散框架 ArtFontDiff; 2) 设计 CFA 与 CCL, 以减轻条件干扰并改善未见组合下的属性敏感生成; 3) 构建更大规模的中英文艺术字数据集 SSAFv2, 为可控艺术字生成研究提供完整的字形一字效组合基准。

1 相关工作

1.1 零样本图像生成

零样本生成关注模型如何将已见类别或条件的知识迁移到未见目标。在生成式零样本学习中,

Chen S 等(2025b)通过特征精炼改善生成质量, Chen S 等(2025a)从相似已见类归纳未见类, Chen S 等(2024)采用进化式架构搜索, Ye 等(2024)引入扩散式数据增强, Chen S 等(2021)则利用层次语义—视觉对齐提升迁移能力。这些研究表明, 零样本泛化不仅取决于生成质量, 也取决于异构条件能否以兼容方式表示与融合。

在图像到图像翻译与可控扩散领域, Pix2Pix(Isola 等, 2017)、FUNIT(Liu 等, 2019)与 StyleGAN 系列(Karras 等, 2019, 2020, 2021)确立了结构保持和风格调制等关键原则, 但在更困难的翻译设定下仍可能不稳定。扩散模型提供了更强的生成先验和更丰富的条件接口。CIA(Alaluf 等, 2024)利用跨图像注意力融合结构与外观; PCDMs(Shen 等, 2023)、IMAGPose(Shen 和 Tang, 2024)和 IMAGDressing-v1(Shen 等, 2025)展示了渐进式或统一式多条件控制在人物图像合成中的有效性。然而, 这些方法主要面向自然图像, 其姿态、服装或外观条件与艺术字所需的离散字符语义和字形拓扑差异较大, 因而只能提供机制启发, 不能直接解决艺术字组合控制生成问题。

1.2 艺术字图像生成

早期艺术字生成方法(Wang 等, 2022)主要分为字效迁移与整体风格迁移。T-Effects(Yang 等, 2017)和 DynTypo(Men 等, 2019)聚焦纹理渲染; TET-GAN(Yang 等, 2019b)、Shape-Matching GAN(Yang 等, 2019a)及 Shape-Matching GAN++(Yang 等, 2022)改进了从普通字符到艺术字效的映射; AGIS-Net(Gao 等, 2019)和 FET-GAN(Li 等, 2020)则将字形与字效作为统一风格联合建模。这些方法能够生成较好的视觉效果, 但结构和渲染仍相互耦合, 控制粒度有限。

为获得更细粒度的结构与风格控制, DSE-Net(Li 等, 2022)、CAFS-GAN(Li 等, 2023)和多因素解耦方法(Zhu 等, 2022)分别对字形与字效因素进行建模。CG-GAN(Kong 等, 2022)进一步引入组件感知模块, 以部件级判别监督增强局部笔画结构和字体风格的一致性; Pan 等(2023)根据目标字符与参考字符的相似性聚合全局风格, 并利用向量量化获得无需人工定义的局部组件表示, 从而兼顾整体字形特征与局部风格细节。这些方法表明, 组件级表征与监督有助于少样本字体生成, 但其主要目标是

从少量字形样本中恢复字体骨架和书写风格,尚未将艺术字效作为独立属性,与字符身份和字形结构共同进行未见组合生成。

近期研究开始利用扩散模型或自回归模型增强字体和风格文字生成(Shu等,2026)。Diff-Font(He等,2024)和FontDiffuser(Yang等,2024)面向单样本字体生成,FontStudio(Mu等,2024)和ArtTypo(Liu等,2025)加强字效或多级控制,Anything to Glyph(Wang等,2023)、DA-Font(Chen W等,2025b)、MX-Font++(Wang等,2025)和FontAnimate(Fu等,2025)则从扩散先验、注意力融合或迁移过程建模等方面改善少样本生成。Pippi等(2025)提出自回归框架,利用变分自编码器和Transformer根据文本内容及风格样例生成任意长度的风格文字,并能够迁移到未见字体或书写风格。然而,该方法侧重文字行级的字体或手写风格模仿,并未分别控制字形结构和艺术字效。区别于上述研究,ArtFontDiff将字符、字形和字效作为三个独立条件,重点研究三类属性在训练阶段未出现的组合下如何协同生成。

2 问题定义

零样本艺术字生成可表示为多条件图像合成问题。给定参考图像三元组 x_c, x_g, x_e ,式中 x_c 表示字符, x_g 表示字形, x_e 表示字效,目标是学习生成映射 $F: x_c, x_g, x_e \rightarrow s$,以合成高质量艺术字图像。形式化地,艺术字图像表示为 $s = Com(c, g, e)$,式中 $Com(\cdot)$ 表示字符 c 、字形 g 与字效 e 的组合。生成图像需要同时满足多重约束:保持 x_c 的语义身份和拓扑结构,继承 x_g 的笔画粗细、连通关系等骨架设计,并呈现 x_e 中的纹理属性和颜色分布。在零样本设定下,推理时使用的具体三元组 (c, g, e) 不出现在训练集中,因此模型需要有效区分这些独立属性,并掌握重新聚合为协调艺术字的组合逻辑。

3 方法

3.1 概述

提出多条件扩散框架ArtFontDiff,如图1所示。模型由三个编码器与条件U-Net主干组成。编码器从字符、字形与字效参考图像 x_c, x_g, x_e 中提取多尺度

潜特征,并在去噪过程中作为引导条件注入U-Net。

该框架主要包含三个关键模块:跨域特征聚合模块(CFA)用于解决异构的字符、字形和字效特征的融合干扰问题,使模型能够以字符为锚点融合字形和字效细节;基于AdaIN(Adaptive Instance Normalization)的跳跃连接(ASC)主要解决U-Net编码与解码器之间的多尺度特征丢失问题,在保留多尺度结构特征的同时,将字形和字效中的风格统计信息更平滑地传递到上采样过程;组合式对比学习(CCL)用于解决模型在面对未见属性组合时容易出现结构扭曲与风格偏移问题,通过拉近生成样本与匹配属性组合的距离、拉远其与不匹配组合的距离,提升模型对字符、字形和字效差异的感知与解耦能力。总体而言,通过特征聚合、风格调制和组合监督这三层设计,该框架有效增强了模型在未见属性组合下的结构保真度与风格可控性。

3.2 多条件艺术字生成

在去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)(Ho等,2020)和去噪扩散隐式模型(denoising diffusion implicit model, DDIM)(Song等,2020)的基础上,构建面向艺术字合成的多条件扩散框架。不同于PromptThis(Guo等,2025)、AnimeDiffusion(Cao等,2024)和DreamAnime(Xu等,2025)等常规条件生成方法,该模型将字符 c 、字形 g 和字效 e 作为相互独立的引导信号来刻画艺术字分布。前向过程中,艺术字图像 $x_0 \sim q(x_0 | c, g, e)$ 在 T 个时间步内逐步加噪;按照DDPM,时间步 t 的噪声潜变量可直接采样为 $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon$,式中 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\bar{\alpha}_t$ 为噪声调度超参数。

方法核心在于反向扩散过程,即字符、字形和字效共同引导去噪轨迹。模型 ϵ_θ 通过观察多属性条件来预测注入 x_t 的噪声,去噪步骤为:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, x_c, x_g, x_e) \right) + \sigma_t z(1)$$

式中 $z \sim \mathcal{N}(0, \mathbf{I})$ 。为实现多条件映射,模型从参考图像 x_c, x_g, x_e 中提取多尺度表征。

3.3 跨域特征聚合

艺术字合成涉及异构信号:字符身份与语义/拓扑结构相关,而字形与字效同时包含结构细节和高频外观细节。直接混合这些信号容易使去噪过程受到风格干扰,从而削弱字形保真。为缓解该问题,所

提框架设计跨域特征聚合模块(CFA),以字符特征作为结构锚点,并以更受控的方式注入字形与字效信息。CFA 主要包含SE注意力和PC注意力两个子模块。

3.3.1 SE注意力模块

由于卷积滤波器只能观察局部邻域,复杂字形结构依赖在多次下采样后可能被削弱。因此采用压缩—激励(squeeze-and-excitation, SE)注意力(Hu等,2018),根据全局重要性重新校准字符通道。具

体而言,字符编码器将字符图像编码为多尺度特征 $f_1^c, f_2^c, f_3^c, \dots$,这些特征与字形、字效特征共同嵌入U-Net的不同CFA模块。在每个CFA模块中,字符特征 f_i^c 与前一U-Net层输出 f_i^o 拼接,经卷积得到 f_i' ;全局平均池化将空间信息压缩为通道描述符 W ,并由全连接层和ReLU学习各通道激活,以重新加权字符特征:

$$f_i^{c'} = F_{reweight}(F_{se}(W)f_i^c) \quad (2)$$

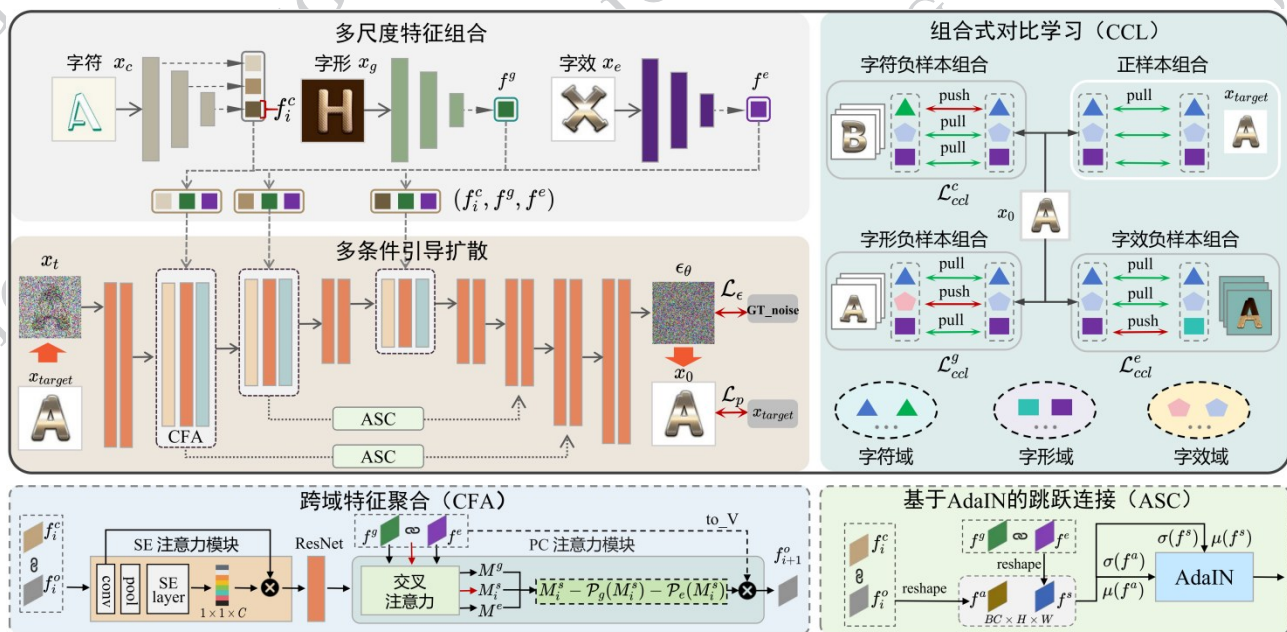


图1 ArtFontDiff多条件组合扩散模型框架

Fig. 1 Framework of the ArtFontDiff multi-condition compositional diffusion model

式中 $F_{se}(\cdot)$ 表示全连接与ReLU层, $F_{reweight}$ 表示基于激活 $F_{se}(W)$ 对特征 f_i^c 进行重加权的过。该步骤用于在多条件融合前重校准字符通道特征,减少结构细节和低频纹理特征对字符骨架的干扰。

3.3.2 PC注意力模块

字形和字效参考仍可能包含无关特征。例如,字形参考可能携带不需要的纹理信息,字效参考也可能包含无关结构。因此,为了尽可能消除这两种风格上的纠缠,提出垂直分量(perpendicular component, PC)注意力机制,其借鉴向量分解思想,通过从联合特征空间中剥离冗余的分量来净化信号。具体而言,首先采用交叉注意力机制计算字形属性注意力图 M^s 与字效属性注意力图 M^e ,以及拼接后的字形-字效联合注意力图 M^s 。字符特征作为查询(Query, Q),对应的字形或字效特征分别作为键

(Key, K)和值(Value, V)。标准交叉注意力图 M 为:

$$M = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (3)$$

该方法将这些注意力图视为高维空间中的基向量,通过减去联合注意力图在单一属性方向上的缩放投影 $\mathcal{P}$,实现特征细化。细化后的注意力图 M' 定义为:

$$M' = M - \mathcal{P}_g(M^s) - \mathcal{P}_e(M^s) \quad (4)$$

$$\mathcal{P}_g(M^s) = \frac{\langle M^s, M^g \rangle}{\gamma \|M^g\|^2} M^g \quad (5)$$

$$\mathcal{P}_e(M^s) = \frac{\langle M^s, M^e \rangle}{\gamma \|M^e\|^2} M^e \quad (6)$$

式中 γ 为缩放系数,用于平衡净化强度。直观上, M'

保留了联合注意力图中与干扰方向 M^g, M^e 对齐程度较低的成分。随后,利用净化后的注意力图对 Value 进行加权求和,从而得到更纯净的融合特征。

3.4 基于 AdaIN 的跳跃连接

为连接编码器和解码器并缓解信息损失,标准 U-Net 常使用跳跃连接。然而,传统跳跃连接缺少将特征调制到目标风格的显式机制,往往难以同时保持笔画保真和颜色一致性。为此,引入基于自适应实例归一化 (adaptive instance normalization, AdaIN) 的跳跃连接 (AdaIN skip connection, ASC), 并依据 Huang 和 Belongie (2017) 的方法进行多尺度风格注入。具体而言,从融合后的字形与字效特征 $f^s = \{f^g, f^e\}$ 中提取均值 $\mu(f^s)$ 和标准差 $\sigma(f^s)$ 等风格统计量,并从 $f^a = \{f_i^e, f_i^o\}$ 中提取内容统计量。每个 ASC 中, $\sigma(f^a)$ 经归一化后由风格统计量调制,再传递到对应上采样层。通过跨尺度动态重校准内容与风格特征,ASC 能在保持清晰结构细节的同时更准确地呈现风格差异。

3.5 组合式对比学习

为缓解未见组合下的结构扭曲和风格偏移,提出组合式对比学习 (CCL)。其目标是提升生成特征对属性不匹配的敏感性,使其更接近目标组合、远离错误组合,增强模型对多属性组合建模的能力。具体来说,CCL 将字符、字形和字效视为可分别对齐的属性维度,通过有针对性地构造单属性错误组合,使模型学习区分不同属性来源的偏差。这种单属性扰动策略有助于模型学习不同属性来源对生成结果的独立影响,避免模型只关注整体图像相似性,而忽略某一属性被错误继承或相互污染的问题。

3.5.1 组合采样策略

令 $s^* = \text{Com}(c^*, g^*, e^*)$ 表示生成结果, $s^+ = \text{Com}(c^+, g^+, e^+)$ 表示所有属性正确对齐的真实目标。为促进稳健解耦,构造三类不匹配组合,每类仅包含一个与 s^+ 相比不匹配属性,如:字符属性不匹配组合 $s^{c-} = \text{Com}(c^-, g^+, e^+)$; 字形属性不匹配组合 $s^{g-} = \text{Com}(c^+, g^-, e^+)$; 字效属性不匹配组合 $s^{e-} = \text{Com}(c^+, g^+, e^-)$ 。训练过程中每类不匹配组合采样 4 个样本。

3.5.2 特征提取与表示

利用字符、字形和字效编码器,将不匹配组合映射到潜空间,得到负特征集合 F^{c-} 、 F^{g-} 和 F^{e-} 。为保

持正属性表征稳定,对错误组合集合中重复出现的正属性取平均,得到中心特征 $\{\bar{f}^{c+}, \bar{f}^{g+}, \bar{f}^{e+}\}$ 。正确目标和生成结果提取的特征分别记为 $\{f^{c+}, f^{g+}, f^{e+}\}$ 与 $\{f^{c*}, f^{g*}, f^{e*}\}$ 。

3.5.3 目标函数

采用 InfoNCE 损失 (He 等, 2020) 衡量生成特征与多属性条件之间的对齐程度:

$$I(\cdot) = -\log \frac{\exp(f^* \cdot f^+ / \tau)}{\exp(f^* \cdot f^+ / \tau) + \sum_{i=1}^n \exp(f^* \cdot f_i^- / \tau)} \quad (7)$$

以字符属性为例,通过直接正特征与平均正特征共同计算字符专属 CCL 损失:

$$\mathcal{L}_{ccl}^c = I(f^{c*}, f^{c+}, F^{c-}) + I(f^{c*}, \bar{f}^{c+}, F^{c-}) \quad (8)$$

总 CCL 损失 $\mathcal{L}_{ccl}$ 由字符、字形和字效三个对比损失相加得到。

3.6 损失函数

该方法包含三类损失函数:去噪损失 $\mathcal{L}_\epsilon$ 、感知损失 $\mathcal{L}_p$ 和组合式对比学习损失 $\mathcal{L}_{ccl}$ 。最终损失:

$$\mathcal{L}_{total} = \lambda_\epsilon \mathcal{L}_\epsilon + \lambda_p \mathcal{L}_p + \mathcal{L}_{ccl} \quad (9)$$

式中 λ_ϵ 和 λ_p 为超参数,用于平衡三类损失函数的权重。去噪损失采用 MSE,使预测噪声逼近真实噪声:

$$\mathcal{L}_\epsilon = \|\epsilon - \epsilon_\theta(x_t, t, x_c, x_g, x_e)\|_2 \quad (10)$$

感知损失 $\mathcal{L}_p$ 用于衡量图像在特征层面的差异:

$$\mathcal{L}_p = \sum_{l=1}^L \|VGG_l(x_0) - VGG_l(x_{target})\| \quad (11)$$

式中 $VGG_l(\cdot)$ 表示视觉几何组网络 (Visual Geometry Group, VGG) (Simonyan 和 Zisserman, 2014) 第 l 层特征, L 为选定层数。组合式对比损失由字符、字形和字效对比损失组成:

$$\mathcal{L}_{ccl} = \lambda_c \mathcal{L}_{ccl}^c + \lambda_g \mathcal{L}_{ccl}^g + \lambda_e \mathcal{L}_{ccl}^e \quad (12)$$

式中 λ_c 、 λ_g 和 λ_e 为超参数。通过上述多项损失函数的联合优化,模型同时受到像素噪声、感知特征和组合属性三个层面的约束。

4 数据集

模型在两个已有艺术字数据集 TE141K (Yang 等, 2021) 和 Fonts (Ge 等, 2020) 上进行评估,并另行构建 SSAFv2,以更充分地验证未见组合的合成能力。表 1 给出了这些数据集的统计比较。

表 1 现有艺术字数据集比较

Table 1 Comparison of existing artistic font datasets

数据集	图像数	字符数	分辨率	字效来源	字形类别	字形-字效组合	语言
Fonts (Ge 等, 2020)	520k	52	128×128	脚本	SP	完整	EN
AGIS-Net (Gao 等, 2019)	1.571M	639	64×64	脚本	SP; CP; HW; CG	无	CN
TE141K-CN (Yang 等, 2021)	50k	775	320×320	设计师	SP	无	CN
TE141K-EN (Yang 等, 2021)	66k	52	320×320	设计师	SP	完整	EN
SSAFv1 (Li 等, 2022)	102k	1,024	320×320	脚本+设计师	SP	完整	EN, CN
SSAFv2	921k	1,024	320×320	设计师	SP; CP; HW; CG	完整	EN, CN



图 2 SSAFv2数据集字形、字效和组合关系概览

Fig. 2 Overview of glyph, effect, and combination relationships in SSAFv2

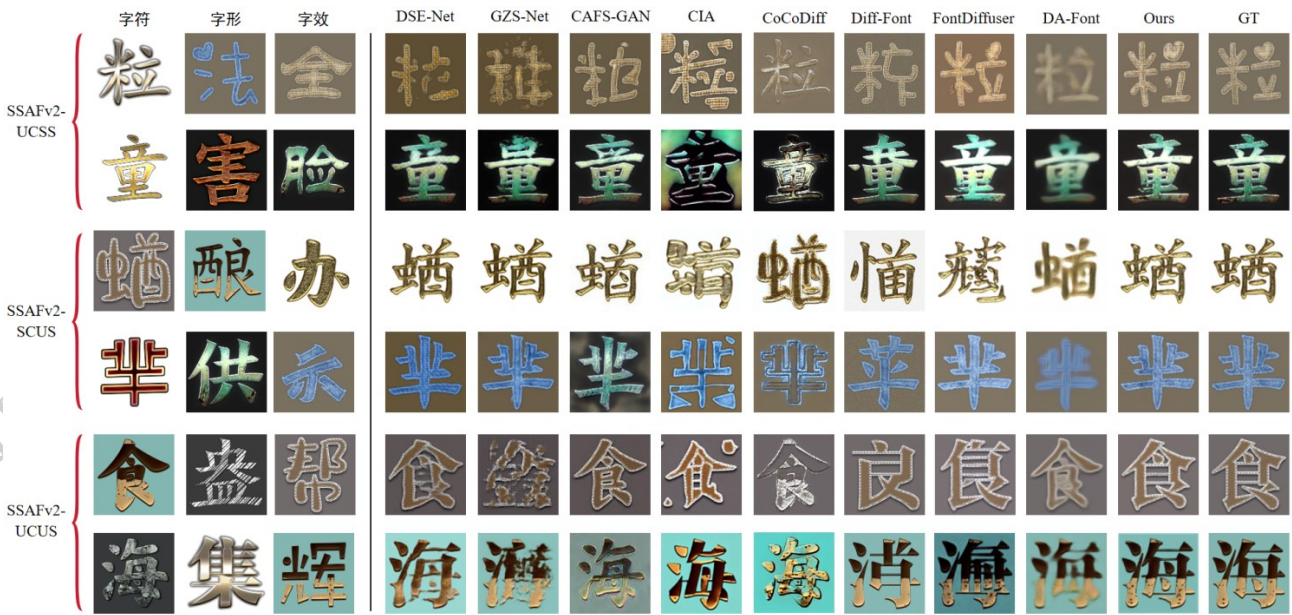


图 3 SSAFv2数据集上的中文艺术字生成结果对比

Fig. 3 Qualitative comparison of Chinese artistic font generation on SSAFv2

4.1 SSAFv2数据集

4.1.1 数据集概述

为弥补现有艺术字数据集在字形多样性、视觉字效复杂度及属性组合完整性方面的不足,本文构

建了SSAFv2数据集。该数据集包含921,600幅分辨率为320×320的艺术字图像,由30种字形和30种字效两两组合形成900种字形一字效类别。字形库包含10种标准印刷体(standard printing, SP)、10种

创意印刷体(creative printing, CP)、5种手写体(handwriting, HW)和5种书法体(calligraphy, CG),覆盖从规则印刷结构到自由书写结构的多种形态。字效库包含30种精细视觉效果,既包括岩石、火焰等自然元素,也包括布料、金属等工业材料,并涵盖纹理、渐变、浮雕、阴影及立体渲染等多种外观表现。SSAFv2共包含52个英文字符和972个中文字符。为扩大中文字符的笔画及结构覆盖范围,本文采用Optset中的775个中文字符,并在此基础上补充197个常用中文字符,所选字符的笔画数覆盖1~24画。由此,字符、字形和字效三类属性形成921,600个字符—字形—字效三元组。其中,每个字符对应900种字形—字效组合,每种字形和每种字效分别对应30,720幅图像,每个字形—字效组合均包含全部1,024个字符。因此,SSAFv2为每一种可能的属性三元组提供唯一对应的艺术字图像,在字符、字形和字效类别内部保持均衡分布,不存在因组合缺失造成的类别长尾问题。完整的属性组合关系也为模型学习字符、字形与字效之间的组合规律提供了充分的数据基础。

4.1.2 数据集制作

SSAFv2的构建过程主要包括基础字形收集、艺术字效设计以及自动化合成与校正3个阶段。

基础字形收集与模板生成。本文从方正字库和iFonts等字体资源平台中选取30种具有代表性的免费TTF字体文件,并按照结构特征将其划分为标准印刷体、创意印刷体、手写体和书法体4类。随后,分别使用这些字体渲染全部中英文字符,并将字符统一生成于320×320分辨率的画布中,作为后续艺术字合成的结构模板。通过引入不同类别的字体,数据集能够覆盖规则笔画、夸张变形、自由书写及书法连笔等多种结构形态。

艺术字效设计。本文使用Adobe Photoshop人工设计30种艺术字效模板。具体而言,通过图案叠加生成不同材质的表面纹理,利用斜面和浮雕操作增强字符边缘的深度感,并结合阴影、渐变、图层置换等操作构建立体视觉效果。所设计的字效包括金属、岩石、布料、饼干和渐变火焰等多种类型,使数据集同时覆盖颜色变化、材质纹理、空间深度和背景装饰等视觉属性。与仅包含简单颜色变化的字效相比,这些复杂纹理能够更充分地检验模型对高频外观细节的提取与迁移能力。

自动化合成与位置校正。在完成字形模板和字效模板设计后,本文利用Photoshop脚本对30种字形和30种字效进行批量组合,生成900种字形—字效模板,并将每种模板应用于全部1,024个字符。为避免不同字符因字面大小、重心位置和笔画分布差异产生明显的空间偏移,本文进一步对生成图像进行逐字符居中与对齐调整,以保证不同属性组合下字符位置的一致性。最终,每个字符、字形和字效三元组均对应一幅独立的艺术字图像。

4.1.3 数据集划分

为在多样条件下评估零样本泛化能力,SSAFv2按训练阶段属性可见性划分为三个子集。UCSS(Unseen Character, Seen Style)用于评估模型对新字符的泛化能力,同时保持已见字形-字效组合;训练使用775个中文字符和44个英文字符,测试使用197个中文字符和8个英文字符。中文字符选取参考Optset(Lian等,2016),并补充常用字符以扩大笔画覆盖范围。SCUS(Seen Character, Unseen Style)评估在已见字符上合成新字形-字效组合,已见与未见字形—字效组合数量之比为9:1。UCUS(Unseen Character, Unseen Style)是最具挑战的设定,其中字符与字形-字效组合在训练阶段均未出现。

4.2 现有艺术字数据集

TE141K-EN是TE141K数据集中的英文子集,包含由设计师制作的英文艺术字样本,具有较高的原始分辨率和完整的英文字符覆盖。该数据集中的样本通常包含较丰富的纹理、颜色和装饰效果,能够反映真实设计场景中英文艺术字的视觉复杂性。TE141K-EN适合用于评估模型在英文艺术字生成中的外观保真能力,尤其是对纹理细节、颜色分布和整体视觉质量的保持能力。Fonts数据集则主要由标准英文字符和字形样式合成得到,包含大量结构较规整的英文样本。与TE141K-EN相比,Fonts的字形结构更规则、视觉字效相对简单,但其样本规模较大,字符覆盖完整,适合检验模型在标准字形条件下的结构保持和跨字形泛化能力。

5 实验

5.1 实现细节

所有实验均在单张NVIDIA RTX 3090图形处理器(graphics processing unit, GPU)上完成。为保证可

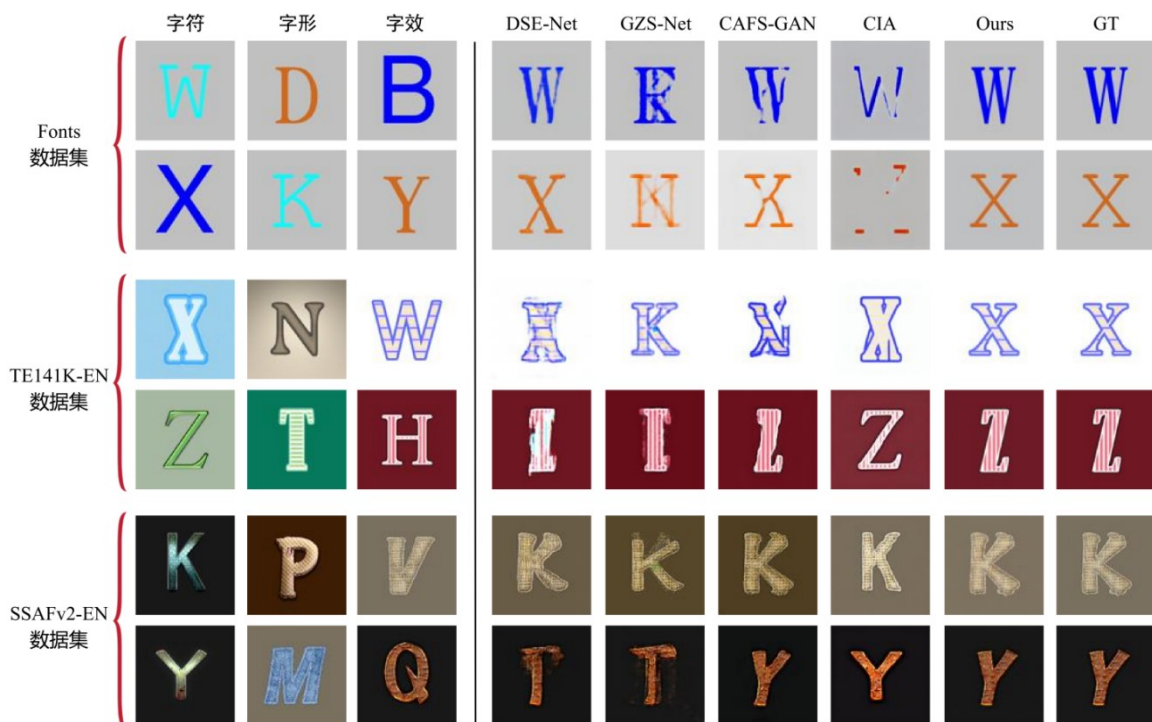


图4 英文艺术字生成结果性对比

Fig. 4 Qualitative comparison of English artistic font generation

比性,在统一适配评测协议下,所有方法均以 96×96 分辨率评估。该选择与可直接复现的开源实现,尤其是 FontDiffuser(Yang 等, 2024) 等扩散基线保持一致,并控制计算开销。尽管源数据集原始分辨率可达 320×320 ,当前协议主要验证共享分辨率下的组合可控性与相对保真。模型训练 500 000 次迭代,批量大小为 16,学习率为 1×10^{-4} 。CCL 每类错误组合采样 4 个样本;超参数 λ_e 、 λ_p 、 λ_c 、 λ_g 和 λ_e 分别设为 1、0.01、0.01、0.01 和 0.01。训练前 400,000 次迭代中,将 λ_c 、 λ_g 和 λ_e 置 0,以避免早期编码器提取无效正负特征。PC 注意力模块中 γ 的值设为 10。训练结束的采样阶段的无分类器引导(classifier-free guidance, CFG)系数设为 7.5,采样算法为 DPM-Solver++, 采样步数为 30 步;该设置下 RTX 3090 生成单张图像约需 1.03 s。

5.2 评价指标

采用 L1 loss、Fréchet inception distance (FID) 和 structural similarity index measure (SSIM) 三个通用指标定量评估 ArtFontDiff 及基线方法。L1 loss 衡量生成图像与真实图像之间的像素级平均绝对误差,结果越小越好;FID 衡量生成图像集合与真实图像集合在特征空间中的分布距离,结果越小越好;SSIM

评估亮度、对比度和结构一致性,值越接近 1 表示结构保持和目标对齐越好。

5.3 对比实验

5.3.1 基线方法

实验将 ArtFontDiff 与多种代表性方法比较,包括支持三个参考样本输入的 DSE-Net(Li 等, 2022)、CAFS-GAN(Li 等, 2023) 和 GZS-Net(Ge 等, 2020) 等零样本艺术字生成模型,通用零样本图像翻译方法 CIA(Alaluf 等, 2024) 和 CoCoDiff(Nie 等, 2026),以及 Diff-Font(He 等, 2024)、FontDiffuser(Yang 等, 2024) 和 DA-Font(Chen 等, 2025) 三种中文字体合成框架。所有基线结果均基于公开的官方 GitHub 实现获得。由于部分基线原生不支持同时输入三张参考图,实验采用统一适配协议:所有方法共享同一字符图像输入;对 CIA、Diff-Font 和 FontDiffuser 等仅支持两张参考图输入的方法,先迁移字形风格,再迁移字效风格。该适配不改变各方法网络结构,而是通过相同训练数据和测试输入实现公平的跨方法比较。

5.3.2 定量实验

在三个 SSAFv2 子集上比较各方法,结果如表 2 所示。ArtFontDiff 在报告指标上整体取得最优结果。GAN 模型如 CAFS-GAN,以及字形专用扩散模型如 FontDiffuser,在较简单设定中具有一定竞争

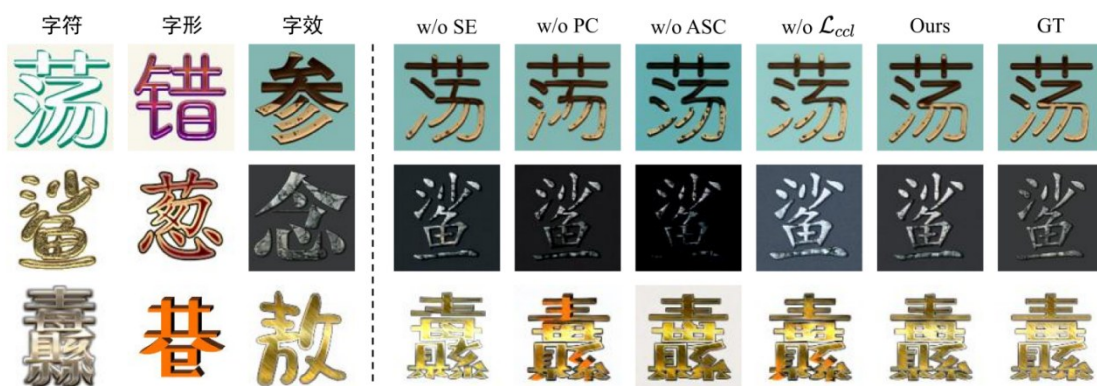


图5 消融实验的定性对比

Fig. 5 Qualitative comparison of ablation experiments

力,但在SSAFv2-SCUS和SSAFv2-UCUS等更困难子集中退化更明显,说明未见字形-字效组合会显著增加属性协调难度。相比之下,ArtFontDiff在未见组合下保持更稳定的定量表现。英文艺术字生成中,SSAFv2-EN、Fonts和TE141K-EN上也呈现类似趋势(表3),说明该框架可迁移到中英文两类艺术字生成任务。

5.3.3 定性实验

图3展示了SSAFv2三个子集上的可视化比较。在SSAFv2-UCSS子集中,由于未见中文字符结构较复杂,现有方法常出现结构坍塌或笔画缺失。例如,部分基线难以重构带有心形结构元素的特殊字形。在SSAFv2-SCUS子集中,已见字符的可读性有所提升,但现有方法仍难以保持字效保真,常表现为颜色扩散或细粒度纹理丢失。实验表明,ArtFontDiff在三个子集中均能保持更清晰的结构对齐和更准确的外观细节。图4进一步展示了英文数据集上的结果。尽管Fonts和TE141K-EN主要由纹理较简单的标准印刷字形构成,现有方法在未见字形-字效组合上仍存在泛化困难;相比之下,ArtFontDiff能更好地保留全局结构与局部风格细节,包括更复杂的创意字形、书法形态和视觉字效。

5.4 消融实验

为系统评估各模块和损失函数的贡献,实验在SSAFv2-UCUS数据集上进行消融实验。消融实验随机选取UCUS测试集中10种字形、10种字效风格和100个字符组成子集。定量和定性结果分别见表4和图5。

SE注意力的有效性:卷积滤波器受局部感受野限制,当复杂字符结构分散到不同通道后,模型可能

难以捕获全局拓扑依赖。如图5所示,当以普通ResNet块替换SE注意力(w/o SE)时,生成结果出现笔画缺失和结构变形;相比之下,所提方法能有效校准通道级字符特征,并通过捕获空间依赖保持艺术字骨架的完整性。

PC注意力的有效性:艺术字参考常存在风格纠缠,例如字形参考可能包含特定纹理干扰。去除PC注意力后(w/o PC),模型交叉注意力图会受到参考样本中无关信息的污染。图5中,字形参考里的橙色立方体字效错误传播到最终输出,形成错误组合。使用PC注意力后,模型可得到更干净的注意力图,并更准确地感知风格及组合形态。

ASC的有效性:U-Net瓶颈中的特征压缩常导致高频结构与风格细节丢失。图5中w/o ASC的结果在前景和背景中均出现明显伪影,笔画边缘也更模糊;ASC则通过跳跃连接过程中的多尺度风格调制,使模型在保留细粒度笔画细节的同时,重构更真实的纹理与颜色组合。

CCL的有效性:异构属性的零样本融合常存在联合兼容性问题,尤其在未见组合中更为突出。去除CCL后(w/o CCL),生成结果出现结构偏移和风格不一致。通过对比正确属性与不匹配属性,CCL提升了模型对属性错误的敏感性,并帮助模型在评测分布内生成更协调的组合。

5.5 组合式对比学习的分析

为进一步验证CCL的有效性,将其与其他对比学习范式比较。表5将CCL与常规实例级对比学习(instance-level contrastive learning, CL)(Chen等, 2020)比较,后者以数据增强版本为正样本、其他样本为负样本;同时分析去除字符、字形或字效对比损

表 2 SSAFv2数据集上中文艺术字生成结果定量比较

Table 2 Quantitative comparison of Chinese artistic font generation on SSAFv2

方法	SSAFv2子集	L1 loss ↓	FID ↓	SSIM ↑
DSE-Net(Li 等, 2022)	UCSS	0.1905	72.11	0.5446
	SCUS	0.2389	37.58	0.7749
	UCUS	0.2580	96.93	0.4509
CAFS-GAN(Li 等, 2023)	UCSS	0.2628	99.58	0.6093
	SCUS	0.2926	198.24	0.4025
	UCUS	0.2910	193.69	0.4111
GZS-Net(Ge 等, 2020)	UCSS	0.3135	61.46	0.3277
	SCUS	0.2027	32.88	0.8574
	UCUS	0.3064	110.29	0.3444
CIA(Alaluf 等, 2024)	UCSS	0.2643	172.79	0.3197
	SCUS	0.2782	69.27	0.3071
	UCUS	0.2849	188.45	0.3159
CoCoDiff(Nie 等, 2026)	UCSS	0.2069	169.74	0.4077
	SCUS	0.2359	145.60	0.2314
	UCUS	0.2715	199.69	0.2667
Diff-Font(He 等, 2024)	UCSS	0.1794	79.09	0.2488
	SCUS	0.1955	94.07	0.2520
	UCUS	0.3203	80.89	0.1752
FontDiffuser(Yang 等, 2024)	UCSS	0.1815	71.37	0.6032
	SCUS	0.1710	63.72	0.4003
	UCUS	0.3284	63.08	0.3487
DA-Font(Chen 等, 2025)	UCSS	0.2597	209.49	0.4904
	SCUS	0.2889	153.16	0.4523
	UCUS	0.2684	188.61	0.3986
Ours	UCSS	0.1754	34.58	0.6351
	SCUS	0.1417	30.19	0.8600
	UCUS	0.1759	52.73	0.4870

注:粗体表示最优值。

失的 CCL 变体。结果表明,相比 CL,所提组合式对比学习更适合艺术字组合任务。

为定性分析属性敏感性,分析使用 ArtFontDiff 的三个编码器从 10 类艺术字中提取特征,并采用降维方法进行可视化,如图 6 所示。在字效特征分析中,以艺术字图像“屋”为锚点,定义正样本组(相同字形和字效)、风格相似组(相同字效但不同字形)和干扰组(不同字效但相同字形)。可视化结果显示,

共享相同字效的艺术字自然聚集;w/o CCL 已表现出粗粒度分组,而完整模型进一步增强了字效特征的类内紧凑性。

对于字形特征,CCL 同样改善了潜空间中结构特征的分离。最明显的提升出现在字符特征中:w/o CCL 版本的字符聚类较为碎片化,仅简单笔画结构被松散聚合;启用 CCL 后,模型能更一致地按语义字符身份组织特征。实验说明 CCL 有助于编码器以更

表 3 英文艺术字生成结果定量比较

Table 3 Quantitative comparison of English artistic font generation

方法	数据集	L1 loss ↓	FID ↓	SSIM ↑
DSE-Net(Li 等, 2022)	Fonts	0.1266	42.75	0.7645
	TE141K-EN	0.1537	56.10	0.6457
	SSAFv2-EN	0.1867	127.88	0.6054
CAFS-GAN(Li 等, 2023)	Fonts	0.2525	45.09	0.8380
	TE141K-EN	0.0994	44.71	0.7671
	SSAFv2-EN	0.2724	111.08	0.5718
GZS-Net(Ge 等, 2020)	Fonts	0.1569	69.22	0.7859
	TE141K-EN	0.1224	78.14	0.7234
	SSAFv2-EN	0.1545	154.48	0.6360
CIA(Alaluf 等, 2024)	Fonts	0.1319	110.08	0.6871
	TE141K-EN	0.1882	116.09	0.6267
	SSAFv2-EN	0.1911	157.12	0.7015
Ours	Fonts	0.1030	11.37	0.8897
	TE141K-EN	0.0753	24.21	0.9194
	SSAFv2-EN	0.0380	62.27	0.9101

注:粗体表示最优值。

表 4 消融实验定量比较

Table 4 Quantitative comparison of ablation experiments

方法	L1 loss ↓	FID ↓	SSIM ↑
w/o SE	0.1578	63.07	0.6750
w/o PC	0.1664	64.59	0.6533
w/o ASC	0.1653	64.14	0.6652
w/o CCL	0.1511	63.38	0.6752
Ours	0.1409	62.01	0.6855

注:粗体表示最优值。

表 5 组合式对比学习各组成部分的消融分析

Table 5 Ablation analysis of compositional contrastive learning components

方法	L1 loss ↓	FID ↓	SSIM ↑
+ CL	0.1514	62.86	0.6543
w/o $\mathcal{L}_{ccl}^c$	0.1524	67.46	0.6295
w/o $\mathcal{L}_{ccl}^g$	0.1532	64.79	0.6416
w/o $\mathcal{L}_{ccl}^e$	0.1570	66.44	0.6232
Ours	0.1409	62.01	0.6855

注:粗体表示最优值。

具属性意识的方式组织特征。

进一步通过对比得分量化多属性感知,使用同一锚点图像“屋”以及相应正样本和干扰集合,计算潜表示之间的余弦相似度。字符对比得分如图 7(a)所示,艺术字图像上方颜色条的灰度越深说明对比分数越高,正样本集合在红框区域相似度最高,而干扰组相似度较低。风格相差越大,对比分数越低;字形与字效得分在图 7(b)和(c)中也呈现相似趋势。

5.6 用户研究

为进一步评估生成艺术字的主观感知质量,进一步开展用户研究。每个测试样例中,不同方法的结果被随机打乱并匿名展示给 20 名参与者。参与者包括具备数字媒体设计经验的 10 名本科生和研究生,以及不具备图形设计、深度学习或计算机相关开发背景的 10 名本科生和研究生;所有参与者均熟悉中英文字符。参与者按 1~5 分从四个维度评分: User-C(字符正确性)、User-T(字形骨架相似度)、User-E(字效相似度)和 User-Q(整体图像质量与主观视觉满意度)。表 6 显示,在该评测设定下,Art-FontDiff 在中英文数据上均获得最高平均评分,为所

表6 用户评估实验结果

Table 6 User study results

方法	中文 User-C	中文 User-E	中文 User-T	中文 User-Q	英文 User- C	英文 User- E	英文 User- T	英文 User- Q
DSE-Net(Li 等, 2022)	3.84	3.16	3.22	3.42	3.98	3.78	4.16	3.52
GZS-Net(Ge 等, 2020)	3.10	2.76	3.36	2.62	3.16	3.10	3.04	3.10
CAFS-GAN(Li 等, 2023)	3.50	3.16	3.22	3.44	3.44	2.24	3.16	3.24
CIA(Alaluf 等, 2024)	3.44	3.44	3.02	3.32	3.50	3.44	2.68	3.38
Diff-Font(He 等, 2024)	2.70	3.88	3.38	3.78	—	—	—	—
FontDiffuser(Yang 等, 2024)	3.96	3.98	3.70	4.24	—	—	—	—
Ours	4.42	4.48	4.50	4.36	4.78	4.78	4.75	4.66

注:粗体表示最优值。

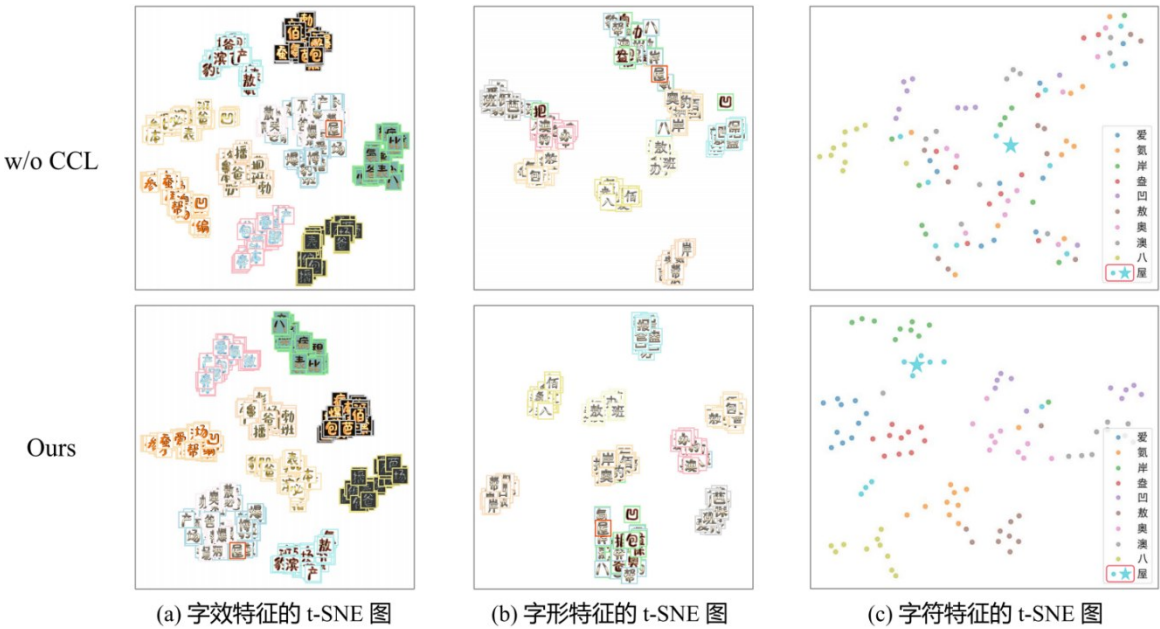


图6 组合式对比学习对艺术字多维属性特征的影响

Fig. 6 Effect of compositional contrastive learning on multidimensional artistic-font features

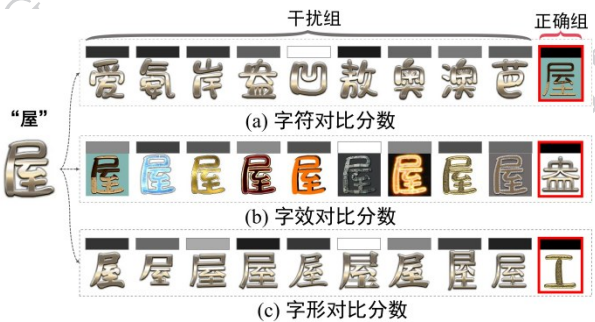


图7 组合式对比学习的多维属性得分

Fig. 7 Multidimensional attribute scores of compositional contrastive learning

提方法在多条件艺术字生成中的有效性提供了主观评价依据。

6 结 论

针对独立字符、字形与字效条件在未见组合中易相互干扰的问题,提出 ArtFontDiff,一个用于可控零样本艺术字生成的多条件扩散框架。该方法结合跨域特征聚合与组合式对比学习,更好地协同字符、字形与字效条件;并构建了 SSAFv2 数据集,将组合设定扩展到更大规模的中英文双语基准。实验证明,ArtFontDiff 在定量与定性研究中均优于对比方

法,并在未见组合上表现出更强鲁棒性。但对多模态条件的可控生成能力仍需进一步验证,后续将研究更丰富条件下的多属性艺术字组合生成。

参考文献

- Alaluf Y, Garibi D, Patashnik O, Averbuch-Elor H and Cohen-Or D. 2024. Cross-Image Attention for Zero-Shot Appearance Transfer//ACM SIGGRAPH 2024 Conference Papers. New York: Association for Computing Machinery: 132: 1-132: 12 [DOI: 10.1145/3641519.3657423]
- Cao Y, Meng X, Mok P Y, Lee T Y, Liu X and Li P. 2024. AnimeDiffusion: Anime Diffusion Colorization. IEEE Transactions on Visualization and Computer Graphics, 30(10): 6956-6969 [DOI: 10.1109/TVCG.2024.3357568]
- Chen S, Chen S, Hou W, Ding W and You X. 2024. EGANS: Evolutionary Generative Adversarial Network Search for Zero-Shot Learning. IEEE Transactions on Evolutionary Computation, 28(3): 582-596 [DOI: 10.1109/TEVC.2023.3307245]
- Chen S, Fu D, Khan S and Khan F S. 2025a. GenZSL: Generative Zero-Shot Learning Via Inductive Variational Autoencoder [EB/OL]. arXiv:2505.11882 [2026-08-01].
<https://arxiv.org/abs/2505.11882> [DOI: 10.48550/arXiv. 2505.11882]
- Chen S, Hong Z, You X and Shao L. 2025b. Semantics-Conditioned Generative Zero-Shot Learning via Feature Refinement. International Journal of Computer Vision, 133(7): 4504-4521 [DOI: 10.1007/s11263-025-02394-7]
- Chen S, Xie G S, Liu Y, Peng Q, Sun B, Li H, et al. 2021. HSVA: Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning [EB/OL].arXiv:2109.15163[2026-08-01].
<https://arxiv.org/abs/2109.15163> [DOI: 10.48550/arXiv. 2109.15163]
- Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A Simple Framework for Contrastive Learning of Visual Representations [EB/OL]. arXiv:2002.05709[2026-08-01].
<https://arxiv.org/abs/2002.05709> [DOI: 10.48550/arXiv. 2002.05709]
- Chen W, Zhu G, Li Y, Ji Y and Liu C. 2025a. Advancements in Chinese Font Generation Since Deep Learning Era: A Survey [EB/OL]. arXiv:2508.06900 [2026-08-01].
<https://arxiv.org/abs/2508.06900> [DOI: 10.48550/arXiv. 2508.06900]
- Chen W, Zhu G, Li Y, Ji Y and Liu C. 2025b. DA-Font: Few-Shot Font Generation via Dual-Attention Hybrid Integration//Proceedings of the 33rd ACM International Conference on Multimedia. New York: Association for Computing Machinery: 6644-6653 [DOI: 10.1145/3746027.3755056]
- Fu B, Wang Z, Yan K, Zhao S, Qin Q, Wen J, et al. 2025. FontAnimate: High Quality Few-Shot Font Generation via Animating Font Transfer Process//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 16015-16025 [DOI: 10.1109/ICCV51701.2025.01486]
- Gao Y, Guo Y, Lian Z, Tang Y and Xiao J. 2019. Artistic Glyph Image Synthesis via One-Stage Few-Shot Learning. ACM Transactions on Graphics (TOG), 38(6): 185: 1-185: 12 [DOI: 10.1145/3355089.3356574]
- Ge Y, Abu-El-Hajja S, Xin G and Itti L. 2020. Zero-shot Synthesis with Group-Supervised Learning[EB/OL]. arXiv:2009.06586 [2026-08-01].
<https://arxiv.org/abs/2009.06586> [DOI: 10.48550/arXiv. 2009.06586]
- Guo Y, Shao H, Liu C, Xu K and Yuan X. 2025. PromptThis: Visualizing the Process and Influence of Prompt Editing During Text-to-Image Creation. IEEE Transactions on Visualization and Computer Graphics, 31(9): 4547-4559 [DOI: 10.1109/TVCG. 2024.3408255]
- He H, Chen X, Wang C, Liu J, Du B, Tao D, et al. 2024. Diff-Font: Diffusion Model for Robust One-Shot Font Generation. International Journal of Computer Vision, 132(11): 5372-5386 [DOI: 10.1007/s11263-024-02137-0]
- He K, Fan H, Wu Y, Xie S and Girshick R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 9729-9738 [DOI: 10.1109/CVPR42600.2020.00975]
- Ho J, Jain A and Abbeel P. 2020. Denoising Diffusion Probabilistic Models[EB/OL]. arXiv:2006.11239 [2026-08-01].
<https://arxiv.org/abs/2006.11239> [DOI: 10.48550/arXiv. 2006.11239]
- Hu J, Shen L and Sun G. 2018. Squeeze-and-Excitation Networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]
- Huang X and Belongie S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE: 1501-1510 [DOI: 10.1109/ICCV.2017.167]
- Isola P, Zhu J Y, Zhou T and Efros A A. 2017. Image-to-Image Translation with Conditional Adversarial Networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 1125-1134 [DOI: 10.1109/CVPR. 2017.632]
- Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, et al. 2021. Alias-Free Generative Adversarial Networks [EB/OL]. arXiv:2106.12423 [2026-08-01].
<https://arxiv.org/abs/2106.12423> [DOI: 10.48550/arXiv. 2106.

- 12423]
- Karras T, Laine S and Aila T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 4401-4410 [DOI: 10.1109/CVPR.2019.00453]
- Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and Improving the Image Quality of StyleGAN//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 8110-8119 [DOI: 10.1109/CVPR42600.2020.00813]
- Kong Y, Luo C, Ma W, Zhu Q, Zhu S, Yuan N, et al. 2022. Look Closer to Supervise Better: One-Shot Font Generation via Component-Based Discriminator//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 13472-13481 [DOI: 10.1109/CVPR52688.2022.01312]
- Li W, He Y, Qi Y, Li Z and Tang Y. 2020. FET-GAN: Font and Effect Transfer via K-Shot Adaptive Instance Normalization//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press: 1717-1724 [DOI: 10.1609/aaai.v34i02.5535]
- Li X, Wu L, Chen X, Meng L and Meng X. 2022. DSE-Net: Artistic Font Image Synthesis via Disentangled Style Encoding//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE: 1-6 [DOI: 10.1109/ICME52920.2022.9859795]
- Li X, Wu L, Wang C, Meng L and Meng X. 2023. Compositional Zero-Shot Artistic Font Synthesis//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. Macao: International Joint Conferences on Artificial Intelligence Organization: 1098-1106 [DOI: 10.24963/ijcai.2023/122]
- Lian Z, Zhao B and Xiao J. 2016. Automatic Generation of Large-Scale Handwriting Fonts via Style Learning//SIGGRAPH Asia 2016 Technical Briefs. New York: Association for Computing Machinery: 12: 1-12:4 [DOI: 10.1145/3005358.3005371]
- Liu K, Wu L, Yu M, Liu X, Xu Y and Meng X. 2025. ArtTypo: Multi-Level Controlled Artistic Typography with Iterative Feedback//Proceedings of 2025 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE: 1-6 [DOI: 10.1109/ICME59968.2025.14208994]
- Liu M Y, Huang X, Mallya A, Karras T, Aila T, Lehtinen J, et al. 2019. Few-Shot Unsupervised Image-to-Image Translation//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 10551-10560 [DOI: 10.1109/ICCV.2019.01065]
- Mao W, Yang S, Shi H, Liu J and Wang Z. 2023. Intelligent Typography: Artistic Text Style Transfer for Complex Texture and Structure. IEEE Transactions on Multimedia, 25: 6485-6498 [DOI: 10.1109/TMM.2022.3209870]
- Men Y, Lian Z, Tang Y and Xiao J. 2019. DynTypo: Example-Based Dynamic Text Effects Transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 5870-5879 [DOI: 10.1109/CVPR.2019.00602]
- Mu X, Chen L, Chen B, Gu S, Bao J, Chen D, et al. 2024. FontStudio: Shape-Adaptive Diffusion Model for Coherent and Consistent Font Effect Generation//Computer Vision - ECCV 2024. Cham: Springer: 305-322 [DOI: 10.1007/978-3-031-73636-0_18]
- Nie W, Li Z, Tao R, Wu B, Wei Y and Zhao Y. 2026. CoCoDiff: Correspondence-Consistent Diffusion Model for Fine-grained Style Transfer[EB/OL].arXiv:2602.14464[2026-08-01].
<https://arxiv.org/abs/2602.14464> [DOI: 10.48550/arXiv.2602.14464]
- Pan W, Zhu A, Zhou X, Iwana B K and Li S. 2023. Few Shot Font Generation Via Transferring Similarity Guided Global Style and Quantization Local Style//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 19449-19459 [DOI: 10.1109/ICCV51070.2023.01787]
- Pippi V, Quattrini F, Cascianelli S, Tonioni A and Cucchiara R. 2025. Zero-Shot Styled Text Image Generation, but Make It Autoregressive//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 7910-7919 [DOI: 10.1109/CVPR52734.2025.00741]
- Shen F and Tang J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation//Advances in Neural Information Processing Systems 37 (NeurIPS). Vancouver: Neural Information Processing Systems Foundation: 6246-6266 [DOI: 10.52202/079017-0202]
- Shen F, Jiang X, He X, Ye H, Wang C, Du X, et al. 2025. IMAGDressing-v1: Customizable Virtual Dressing//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press: 6795-6804 [DOI: 10.1609/aaai.v39i7.32729]
- Shen F, Ye H, Zhang J, Wang C, Han X and Yang W. 2023. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models[EB/OL].arXiv:2310.06313 [2026-08-01].
<https://arxiv.org/abs/2310.06313> [DOI: 10.48550/arXiv.2310.06313]
- Shu Y, Zhao F M, Chen Z Y, Zhao T Q, Wang Y Z, Li K C, et al. 2026. Survey on visual text generation techniques. Journal of Image and Graphics, 31(6): 2103-2124 (舒言, 赵方敏, 陈泽宇, 赵天齐, 王逸竹, 李焜焱, 等. 2026. 可视文本生成技术综述. 中国图象图形学报, 31(6): 2103-2124) [DOI: 10.11834/jig.260047]
- Simonyan K and Zisserman A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition [EB/OL]. arXiv: 1409.1556 [2026-08-01].
<https://arxiv.org/abs/1409.1556> [DOI: 10.48550/arXiv.1409.1556]
- Song J, Meng C and Ermon S. 2020. Denoising Diffusion Implicit Models [EB/OL]. arXiv:2010.02502 [2026-08-01].
<https://arxiv.org/abs/2010.02502> [DOI: 10.48550/arXiv.2010.02502]

- 02502]
- Wang C, Wu L, Liu X, Li X, Meng L and Meng X. 2023. Anything to Glyph: Artistic Font Synthesis via Text-to-Image Diffusion Model//SIGGRAPH Asia 2023 Conference Papers. New York: Association for Computing Machinery: 55: 1-55: 11 [DOI: 10.1145/3610548.3618208]
- Wang C, Wu G H, Yao Y, Ren Y Z, Wang Q H and Yuan L F. 2022. Review of Chinese characters generation and font transfer based on deep learning. *Journal of Image and Graphics*, 27(12): 3415-3428 (王晨, 吴国华, 姚晔, 任一支, 王秋华, 袁理锋. 2022. 深度学习汉字生成与字体风格迁移综述. *中国图象图形学报*, 27(12): 3415-3428) [DOI: 10.11834/jig.210360]
- Wang W, Sun D, Zhang J and Gao L. 2025. MX-Font++: Mixture of Heterogeneous Aggregation Experts for Few-Shot Font Generation//Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE: 1-5 [DOI: 10.1109/ICASSP49660.2025.10888465]
- Xu C, Xu Y, Zhang H, Xu X and He S. 2025. DreamAnime: Learning Style-Identity Textual Disentanglement for Anime and Beyond. *IEEE Transactions on Visualization and Computer Graphics*, 31(8): 4198-4209 [DOI: 10.1109/TVCG.2024.3397712]
- Yang S, Liu J, Lian Z and Guo Z. 2017. Awesome Typography: Statistics-Based Text Effects Transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE: 7464-7473 [DOI: 10.1109/CVPR.2017.308]
- Yang S, Liu J, Wang W and Guo Z. 2019b. TET-GAN: Text Effects Transfer via Stylization and Destylization//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press: 1238-1245 [DOI: 10.1609/aaai.v33i01.33011238]
- Yang S, Wang W and Liu J. 2021. TE141K: Artistic Text Benchmark for Text Effect Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3709-3723 [DOI: 10.1109/TPAMI.2020.2983697]
- Yang S, Wang Z and Liu J. 2022. Shape-Matching GAN++: Scale Controllable Dynamic Artistic Text Style Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7): 3807-3820 [DOI: 10.1109/TPAMI.2021.3055211]
- Yang S, Wang Z, Wang Z, Xu N, Liu J and Guo Z. 2019a. Controllable Artistic Text Style Transfer via Shape-Matching GAN//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE: 4442-4451 [DOI: 10.1109/ICCV.2019.00454]
- Yang Z, Peng D, Kong Y, Zhang Y, Yao C and Jin L. 2024. FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press: 6603-6611 [DOI: 10.1609/aaai.v38i7.28482]
- Ye Z, Gowda S N, Huang X, Xu H, Jin Y, Huang K, et al. 2024. Zero-Diff: Solidified Visual-Semantic Correlation in Zero-Shot Learning [EB/OL]. arXiv:2406.02929 [2026-08-01].
<https://arxiv.org/abs/2406.02929> [DOI: 10.48550/arXiv.2406.02929]
- Zhu A, Yin Z, Iwana B K, Zhou X and Xiong S. 2022. Text Style Transfer Based on Multi-Factor Disentanglement and Mixture//Proceedings of the 30th ACM International Conference on Multimedia. New York: Association for Computing Machinery: 2430-2440 [DOI: 10.1145/3503161.3548239]

作者简介

李祥,男,博士研究生,主要研究方向为计算机视觉,三维生成。E-mail:xiangli_@mail.sdu.edu.cn

孟雷,通信作者,男,教授,主要研究方向为计算机视觉/多媒体计算、机器学习和推荐系统。E-mail:lmeng@sdu.edu.cn

武蕾,通信作者,女,副教授,主要研究方向为多媒体计算和智能内容生成。E-mail:i_lily@sdu.edu.cn

孟祥旭,男,教授,主要研究方向为计算机图形学、图像处理和网格计算。E-mail:mxx@sdu.edu.cn