中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：22721547

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	基于形状空间理论特征增强的小样本图像生成方法研究与应用

作　　者　阮礼恒
学科专业　计算机应用技术
导　　师　韩越兴
完成日期　二〇二五年四月

姓　　名：阮礼恒　　学号：22721547

论文题目：基于形状空间理论特征增强的小样本图像生成方法研究与应用

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主　　席：

委　　员：

导　　师：

答辩日期：2025 年 6 月 10 日

姓　　名：阮礼恒　　学号：22721547

论文题目：基于形状空间理论特征增强的小样本图像生成方法研究与应用

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：阮礼恒

日期：2025 年 6 月 10 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

(保密的论文在解密后应遵守此规定)

学位论文作者签名：阮礼恒　　导师签名：[illegible]

日期：2025 年 6 月 10 日　　日期：2025 年 6 月 10 日

上海大学工学硕士学位论文

基于形状空间理论特征增强的小样本图像生成方法研究与应用

作　　者: 阮礼恒
导　　师: 韩越兴
学科专业: 计算机应用技术

计算机工程与科学学院
上海大学
2025 年 4 月

A Dissertation Submitted to Shanghai University for the
Degree of Master in Engineering

Research and Application of Few-Shot Image Generation Method with Feature Augmentation Based on The Shape Space Theory

Candidate: Liheng Ruan
Supervisor: Yuexing Han
Major: Technology of Computer Application

School of Computer Engineering and Science
Shanghai University
April, 2025

摘　要

以生成对抗网络和扩散模型为代表的深度生成模型，不仅在图像生成领域展现出巨大潜力，也有力地推动了人工智能技术的整体发展。然而，这些模型由于对大规模训练数据的依赖，在处理信息受限问题时会面临性能瓶颈。主要挑战体现在两个方面：一，很多研究和工业领域缺乏大规模数据对模型进行预训练，在没有有效的预训练模型的情况下，需要从零开始，利用少量样本进行训练，深度挖掘少量样本内在分布规律以生成高质量、多样化的图像；二，在零样本文本引导的风格迁移任务中，一些预训练生成模型未曾学习过的风格信息，需要高效地融合到生成过程中，实现生成图像的风格控制。为了应对这些挑战，本论文基于形状空间理论，提出一种预形状空间测地曲面非线性特征增强策略，旨在利用数据的内在结构，深度挖掘小样本场景下稀疏信息，促进在零样本风格迁移中新颖风格信息的高效融合与精确控制，最终提升模型的生成性能。主要研究工作包括：

（1）针对图像生成模型训练时面临的训练样本过少及缺乏适用预训练模型等问题，提出了基于预形状空间测地曲面信息迁移的方法。该方法克服了从零开始训练时难以有效学习极少样本分布的瓶颈，以生成高质量、多样化的图像。该方法首先提取少量样本的深度特征，随后在预形状空间中利用这些特征构建测地曲面，进行非线性特征增强。接着，基于增强后的特征构建伪源域以模拟更丰富的数据分布，并进行从伪源域到目标域的信息迁移。最终，在信息迁移阶段施加插值监督与正则化约束进行优化。实验证明，相较于现有方法，本方法在多领域数据集上，显著提升了生成图像的质量、细节丰富度和多样性，有效缓解了模式坍塌，并展示了其生成图像在辅助下游任务中的潜力。

（2）针对文本引导的零样本图像风格迁移任务的挑战，提出了基于预形状空间中测地曲面特征增强的零样本风格迁移方法，高效地将外部新颖风格信息注入预训练模型，同时确保风格一致性与内容准确性。该方法将测地曲面特征增强思想应用于基于预训练扩散模型的风格迁移框架，结合滑动窗口裁剪处理局部信息，并利用测地曲面特征增强模块在预形状空间中促进风格与内容特征的有效融合。实验表明，该方法能在无需额外模型微调或风格参考的情况下，实现灵活的文本引导风格控制，

并在生成具有目标风格的图像时，相较于对比模型，较好地维持了原内容结构。

本论文探索了将形状空间理论及测地曲面结构用于深度生成模型特征增强的可行性。通过所提出的非线性特征增强策略，利用数据的内在结构信息，针对小样本图像生成和零样本图像风格迁移中的特定挑战提出了相应的解决方案。实验结果表明，该方法在提升生成图像的质量、多样性以及风格保真度方面取得了积极效果，显示了结合形状空间理论来改善数据稀疏环境下生成模型性能的潜力。

关键词：小样本图像生成；图像风格迁移；形状空间理论；特征增强

ABSTRACT

Deep generative models, such as Generative Adversarial Networks and Diffusion Models, have shown great potential in image generation, driving advancements in artificial intelligence. However, their dependence on large-scale training data creates bottlenecks when information is limited. The main challenges are twofold: first, in many research and industrial domains, suitable large-scale datasets or effective pre-trained models are unavailable. Training from scratch with small samples makes it hard to learn the underlying data distribution deeply enough to generate both high-quality and diverse images; second, when applying a new style that the pre-trained model has never seen, it is challenging to integrate the unseen style information efficiently into the generation process while ensuring both accurate stylization and content preservation. To address these issues, this thesis introduces the Shape Space theory and proposes a nonlinear feature augmentation strategy based on Geodesic surfaces in the Pre-Shape Space. By leveraging the intrinsic structure of data, the proposed method both mines sparse information in few-shot settings and fuses novel style information in zero-shot image style transfer, ultimately boosting overall generation performance. The main contributions are as follows:

(1) To address the challenges of insufficient training samples and lack of suitable pre-trained models, an information transfer method based on Geodesic surfaces in the Pre-Shape Space is proposed. First, deep features are extracted from the limited samples. Next, Geodesic surfaces are built in the Pre-Shape Space using these features to achieve nonlinear feature augmentation. A pseudo-source domain is then constructed from the augmented features to simulate a richer data distribution, and information is transferred from the pseudo-source domain to the target domain of generator. Finally, interpolation supervision and distance regularization constraints are applied during transfer to optimize performance. Experimental results on multiple datasets demonstrate that, compared to existing methods, the proposed method enhances image quality, detail richness, and diversity, effectively mitigates mode collapse, and shows potential for supporting downstream tasks.

(2) For the zero-shot text-guided image style transfer task, a style transfer method based on feature augmentation on Geodesic surface in the Pre-Shape Space is introduced to efficiently inject external style information into a pre-trained model while ensuring style consistency and content fidelity. Applying the principle of feature augmentation on the Geodesic surface within a style transfer framework based on pre-trained diffusion models, this method processes local information using sliding window crop and leverages the feature augmentation on Geodesic surface module to facilitate the effective fusion of style and content features in the Pre-Shape space. Experimental findings indicate that this method enables flexible text-guided style control without additional model fine-tuning or style references and better preserves the original content structure compared to baseline methods.

This thesis explores the feasibility of using the Shape Space theory and Geodesic surface for feature augmentation in deep generative models. The proposed nonlinear feature augmentation strategy leverages the intrinsic structural information of data to address specific challenges encountered in few-shot image generation and zero-shot image style transfer. Experimental results indicate that the proposed method achieved positive effects in enhancing the quality, diversity, and style fidelity of generated images, demonstrating the potential of using the Shape Space theory to improve the performance of generative models in data-sparse environments and providing useful references for research in related directions.

Keywords: Few-shot Image Generation; Image Style Transfer; The Shape Space Theory; Feature Augmentation

目 录

第一章　绪论

1.1　课题来源

本课题得到国家自然科学基金（面上，编号：52273228），云南省科技计划重点项目（编号：202302AB080022），上海大学硅酸盐文物保护教育部重点实验室开放课题项目（编号：SCRC2023ZZ07TS）的资助。

1.2　课题背景概述

人工智能（Artificial Intelligence，AI）技术的蓬勃发展及其广泛应用已成为时代趋势，其战略重要性日益凸显。在众多 AI 技术分支中，机器视觉作为实现环境感知和理解的核心技术，对 AI 系统的性能和应用落地起着决定性作用，是当前研究的热点领域。

近年来，以生成对抗网络（Generative Adversarial Networks，GANs）[1] 和扩散模型（Diffusion Models，DMs）[2-3] 为代表的生成模型在计算机视觉领域取得了显著进展。这些模型能够合成具有高逼真度和多样性的图像，并在数据增强、艺术创作、虚拟现实及科学模拟等多个领域获得广泛应用[4-5]。然而，这些强大模型的训练普遍依赖于海量、多样化的数据集，通过学习大规模样本来捕捉复杂的数据分布。模型对大规模训练数据的依赖，在许多实际应用中构成了挑战，诸如新材料研发、医学影像、工业检测和稀有物种记录等领域，其数据获取往往受到高成本、隐私限制或客观条件的制约[6]。

当可用的样本或明确的指导信息极其有限时，现有深度生成模型往往难以有效学习数据的内在结构和变化规律，导致生成结果质量下降、缺乏多样性，甚至出现模式坍塌（Mode Collapse）问题[7]。如何在信息受限条件下应对高质量、可控图像生成的挑战，是当前生成模型领域亟待解决的关键难题之一。这体现在两个典型场景：一是在不依赖任何大规模预训练模型、完全从零开始训练的情况下，需要从极少量样本中学习数据分布并生成新图像的小样本图像生成（Few-shot Image Generation，FSIG），这考验了模型从极端稀疏数据中深度挖掘信息的能力；二是需要在预训练生成模型本身并未学习过目标风格知识的情况下，根据外部信息引导模型生成具有该

全新风格且保持原始内容结构的图像，即零样本风格迁移（Zero-shot Style Transfer）。这两个场景的共性挑战在于都面临极端的信息限制。无论是处理稀疏数据还是融合未见过的外部信息，核心难题都是如何在有限的指导信息下，有效平衡生成图像的保真度、多样性与内容和结构的一致性。

尽管研究者们提出了基于迁移学习[7-8]，以及数据增强或特征增强[9-10]等策略来缓解信息稀疏性，但前者受限于预训练模型的可用性和领域相关性，后者在面对极端少量样本时，简单的增强手段往往不足以捕捉复杂的非线性数据分布或实现细粒度的内容与风格平衡。同样，对于零样本风格迁移，利用大型视觉-语言模型[11]提供文本与图像关联信息的方法[12]，虽然为零样本控制提供了强大潜力，但其主要挑战是将外部模型提供的、描述新颖风格的引导信息，有效转化为对生成模型内部过程的精确、细粒度控制信号。这要求在高效注入新风格的同时，确保内容结构的稳定。这些现有方法的局限性凸显了在信息极端受限条件下，当前技术路线普遍缺乏有效利用数据内在结构特性进行特征增强或信息融合的机制。因此，它们难以仅从少量样本中捕捉真实的分布规律，或在注入新风格时精确控制生成过程并保持内容一致性。

为了突破这些瓶颈，推动生成模型在更广泛的现实场景中落地应用，亟需探索新的理论与方法。一个富有潜力的研究方向是转向深入挖掘和利用数据本身所蕴含的内在结构特性，而不再仅仅依赖增加数据量或改进现有网络结构。借鉴形状空间等理论，通过关注数据的本质结构信息，并利用样本间的非线性关联，在数据和信息受限的挑战性场景下设计出更高效、更鲁棒的生成模型。

1.3 课题研究目的与意义

针对深度生成模型在处理信息受限场景时存在的性能瓶颈，本论文的核心研究目的在于开发并验证一种基于形状空间理论的特征增强新策略。具体而言，本研究旨在通过应用形状空间理论与测地结构设计非线性特征增强算法，并将其有效整合入 GAN 或扩散模型等先进框架，进而构建与评估能够显著改善极端小样本图像生成质量及多样性的方法，同时探索与验证该策略在提升零样本文本引导风格迁移控制精度和内容保持能力方面的应用潜力。最终，通过全面的实验评估，证实所提策略在应对上述信息受限挑战方面的有效性与优越性，证明其作为相关领域实用工具的

价值。

本研究的意义主要体现在其潜在的理论贡献和实际应用价值。在理论层面，本研究的意义在于提出并验证了一种受几何理论启发的特征增强技术。该技术利用预形状空间的测地曲面进行非线性特征增强，为小样本学习和特征工程领域提供了一种不同于传统线性混合或简单变换的实用方法选项。它展示了将形状空间概念作为工具应用于解决现代深度学习挑战的可行性，并为探索在无大规模预训练条件下实现高质量信息受限生成提供了有价值的实践案例和经验。在应用层面，这项工作直接面向数据稀缺性这一普遍存在于材料科学、医学影像、生物信息学等众多领域的痛点，所提出的方法有望为这些领域提供更强大的图像生成与数据增强能力，从而降低对海量标注数据的依赖，辅助科学研究和工程应用。同时，通过提升文本控制风格迁移任务的可控性与内容保真度，研究成果可为计算机辅助设计、个性化内容创作等领域提供更灵活、更高质量的技术支持。

1.4 国内外研究现状

本论文结合基于形状空间理论的特征增强，实现小样本图像生成以及零样本图像风格迁移。本节介绍形状空间理论、特征增强、小样本图像生成以及图像风格迁移的研究现状。

1.4.1 形状空间理论研究现状

流形（Manifold）作为描述数据内在结构的重要数学工具，已在计算机科学的众多分支中得到成功应用[13]。随着几何深度学习（Geometric Deep Learning，GDL）的兴起[14]，利用流形理论理解和处理具有复杂结构的数据成为研究热点。在深度学习中，流形假设有助于揭示高维数据的低维本质结构。例如，在降维与可视化任务中，t-SNE[15] 等方法旨在将数据投影到低维流形。在对比学习的度量方法上，常将特征约束在单位超球面（Unit Hyper-sphere）上[16]。在生成模型领域，理解数据潜在流形的结构特性被认为有助于合成高质量样本[17]，而流形正则化（Manifold Regularization）[18] 等技术也通过引入几何约束来提升模型性能。

形状空间（Shape Space）理论，由 Kendall 在 20 世纪 80 年代系统提出[19]。它聚焦于物体在滤除了平移、旋转和尺度等无关的相似变换影响后所保留的内在几何形态信息。该理论的核心构造是预形状空间（Pre-Shape Space），它通常通过将一组

中心化、尺度归一化的标志点（Landmarks）构型嵌入到一个高维单位超球面上，而形状空间本身则是预形状空间在旋转群作用下的商空间，具有黎曼流形（Riemannian Manifold）的结构。这套理论提供了一个强大的框架，用于形状的表示、比较和变换。

基于其严谨的几何基础，形状空间理论在多个领域得到了应用。早期工作利用形状空间中的测地曲线（Geodesic Curve）及其插值，实现了三维模型间的平滑、自然的变形[20]。在目标识别与匹配任务中，研究者通过将目标轮廓或关键点投影到预形状空间，并计算测地距离或寻找测地路径来进行形状比对与分类[21-22]。近年来，形状空间理论与深度学习的结合日益紧密。例如，Paskin 等人[23] 利用形状空间理论中的先验知识从 2D 图像推断 3D 鲨鱼骨骼姿态，显示了其在数据有限条件下进行 3D 重建的潜力。Friji 等人[24] 则将形状空间理论与等变神经网络（Equivariant Neural Networks）相结合，在人体姿态识别任务中取得了优异表现，验证了显式建模几何不变性对于特定任务的优势。

除识别和姿态估计外，形状空间理论也开始渗透到生成模型和数据增强领域。KS-VAE（Kendall Shape Variational Auto-Encoder）通过设计等变编码器-解码器，将图像信息编码到形状空间潜变量中，成功实现了形状与姿态的解耦，提升了 VAE 的表示能力和可解释性[25]。而在特征增强方面，FAGC 提出在预形状空间中构建连接同类样本特征的测地曲线，并沿曲线采样生成新特征，用于小样本学习[26]。这些工作表明，利用形状空间的几何结构进行表示学习或数据增强具有独特优势，特别是在需要处理姿态变化或样本量有限的场景。

然而，当前形状空间理论与深度学习的融合研究仍面临一些挑战和局限。多数成功应用集中在姿态估计、特定对象如人脸、骨架识别或三维重建等几何结构信息相对明确的任务上。一个能够处理通用图像、将形状空间理论与主流图像生成模型深度融合的系统性框架尚待完善。特别是在通用图像生成任务中，如何将侧重于分析几何形态的形状空间理论，与生成模型需要处理的复杂纹理、光照、背景等外观信息有效结合，仍然是一个关键难题。因此，尽管形状空间理论作为分析形状内在属性的有力工具，其在形状重建、姿态估计和特定数据增强任务中已展示潜力，但在更复杂、更通用的任务，如小样本图像生成中，其潜力仍需进一步挖掘。

1.4.2 特征增强研究现状

数据增强（Data Augmentation）是深度学习中广泛采用以提升模型泛化能力和鲁棒性的关键技术之一，其目标在于通过扩充训练数据的规模和多样性。传统上，最常见的数据增强方法直接在原始输入数据空间进行操作，例如对图像进行随机裁剪、旋转、色彩变换等几何或光度调整。然而，这些输入空间的增强方法也存在局限性。一方面，它们产生的变换可能相对表层，未必能有效模拟数据在高维特征空间中可能出现的深层语义变化。另一方面，在某些特定领域，如材料科学研究或医学影像分析中，过度或不恰当的输入变换甚至可能破坏关键的细微结构信息，或生成脱离真实数据分布的无效样本。此外，这些领域往往面临标注数据获取困难的挑战，使得对增强有效性的需求更为迫切。

为了克服输入空间增强的这些不足，并寻求更直接作用于模型学习到的数据表示的方法，研究者们提出了特征增强（Feature Augmentation）技术。该策略的核心思想是在深度神经网络学习到的高维特征空间中进行数据扩展，而非输入端。通过对特征向量进行混合、插值或其他变换，特征增强旨在生成更具语义意义、更能体现数据内在结构的新特征样本，从而为提升模型性能开辟了新的途径。

早期的特征增强方法主要包括特征噪声注入、特征插值和外插等策略[1,27]。DeVries 和 Taylor[27] 提出，在通过无监督学习获得的特征空间中，应用简单的变换来扩充数据集，例如向特征向量添加随机噪声，或在特征空间中对两个样本的特征向量进行线性插值。这些方法的优势在于其领域无关性，无需针对特定数据类型设计复杂的变换。

随后，数据混合（Data Mixing）成为特征增强中的重要策略。Mixup 通过对成对样本的输入和标签进行凸组合来生成虚拟训练样本，鼓励网络在训练样本之间表现出更简单的线性行为，从而提高模型的泛化能力和对对抗样本的鲁棒性[28]。Manifold Mixup 将 Mixup 的思想扩展到了隐藏层，对该层输出的特征表示以及对应的标签进行凸组合，旨在生成更平滑的多层级决策边界，学习到更扁平化的类表示[10]。类似地，特征增强操作也在迁移学习和多模态场景中被证明有效[29]。

研究者们不断探索更复杂的特征空间操作，以提升增强样本的多样性、语义丰富性与任务相关性。例如，FeatMatch 通过学习类原型并进行特征精炼来生成更符合类别分布的样本[30]。Mangla 等人[31] 结合自监督学习预训练获取更具结构意义的特征

流形，再进行 Manifold Mixup 等增强操作，生成更具合理性的新样本。Khan 等人[32]通过在低层与高层特征空间分别训练生成模型并进行采样，大幅丰富了数据样本的多样性。MixStyle 通过随机混合样本间的如均值与方差等特征统计信息来模拟风格变化以提升域泛化能力[33]。Chu 等人[34] 利用特征级数据增强来改善长尾分布数据集的识别表现，Liu 等人[35] 则将 SMOTE 方法应用于特征空间，进而在故障样本不足的场景下扩充数据，从而提高模型性能。

尽管以上特征增强方法在提升模型性能方面取得了一定的成功，但它们直接在原始特征空间中进行操作。在该空间内，简单的线性插值可能无法有效捕捉数据样本间潜在的结构性关联和有意义的变化模式[27]。真实数据的内在组织方式往往更为复杂，其样本间的合理演变路径并非简单的直线。因此，这种直接在原始空间进行线性增强的方式，可能难以生成完全遵循真实数据内在结构及其合理变化的样本，从而影响其真实感和语义合理性。这表明，探索更能反映数据本质结构的表示空间及适配该空间的特征增强策略可能更有优势。

为了克服这一瓶颈，近期研究的一个重要趋势是将几何感知策略引入特征增强。其中，基于形状空间理论的方法展现了独特的潜力。Han 等人[26] 提出的 FAGC 方法，将在预形状空间中构建测地曲线，并沿此非线性路径进行特征插值与增强，该方法利用 ViT 提取特征，在极少样本情景下显著提升了数据多样性。同时，如 KS-VAE[25] 等工作，通过在 VAE 框架中显式引入 Kendall 形状空间约束，实现了姿态与形状的有效分离，也间接促进了特征增强的有效性。这些基于形状空间等几何理论的特征增强方法，因其能更准确地捕捉数据的内在结构属性，有望生成在语义和结构上更一致、更多样化的样本。

1.4.3 小样本图像生成研究现状

深度生成模型，如变分自编码器（VAE）[36]、生成对抗网络（GAN）[1] 和扩散模型（DM）[2-3]，在生成高保真、多样化图像方面取得了巨大成功。然而，其性能高度依赖大规模训练数据，这在医学影像、材料科学、遥感监测等数据获取受限的领域构成了应用瓶颈[6,37]。因此，如何在数据有限条件下生成高质量、多样化的图像，即小样本图像生成（Few-shot Image Generation，FSIG），并可能将其用于如分类、分割等下游任务的数据增强[37]，已成为一个关键研究方向。现有 FSIG 研究大致可分为两大主流范式：

第一类是基于源域（Source Domain）的迁移方法（Transfer Learning）。这类方法借鉴迁移学习思想，将在大型相关数据集上预训练的生成模型（源域）适配到样本有限的目标域[38]。核心在于利用源域学习到的丰富先验知识来弥补目标域数据的不足。具体技术包括直接微调部分或全部模型参数、引入正则化项以保留源域知识和多样性，如跨域距离一致性损失（Cross-Domain Correspondence，CDC）[7]、结构信息对齐（Relaxed Spatial Structural Alignment，RSSA）[8]、双对比学习（Dual Contrastive Learning，DCL）[39]，或如 MineGAN[40] 以及 LFS-GAN[41]，采用参数隔离与轻量级调制技术提高效率并对抗灾难性遗忘。这类方法在极端小样本条件下通常表现更优，但也面临固有挑战：强依赖于高质量预训练模型和相关源域的可用性，且在保留源先验与适应目标细节之间存在微妙的适配困境（Adaptation Dilemma）[42]，对域差异较大的任务效果有限。

第二类是无源域的方法，这类方法旨在仅利用目标域的少量样本进行训练，避免依赖额外的源域数据或预训练模型。一种常用策略是数据增强，例如可微增强技术（DiffAugment，DA）[9] 和自适应判别器增强（Adaptive Discriminator Augmentation，ADA）[4]，通过在线增强真实和生成样本来缓解判别器的过拟合问题。另一种思路是改进模型结构或训练策略，如 FastGAN 通过引入跳跃连接激励和自监督学习来提升训练效率和稳定性[43]。针对扩散模型，PatchDiffusion 提出了一个基于图像块（Patch）的训练框架，旨在降低扩散模型对大规模数据和训练时间的需求，并改善其在小样本数据集上的表现[44]。针对更极端的数据稀疏场景，SinGAN[45]、CoSinGAN[46] 以及或单样本扩散模型 SinDiffusion[47]，探索了从单张图像学习生成模型的可行性。HP-VAE-GAN 的改进版本则专注于生成如材料图像等特定类型的图像用于数据增强，但其适用性可能受限于图像的纹理特性[37,48]。无源域方法的核心挑战在于，需要直接从极度稀疏的样本中学习复杂的真实数据分布，极易陷入对训练样本的过拟合或在追求多样性时发生模式坍塌。MixDL 在此背景下提出了一种创新方法，它无需额外数据或预训练，通过引入基于 Mixup 的距离正则化策略，在显著提升多样性的同时，较好地保持了图像逼真度，因此被广泛视为该特定场景下的一个关键基线和代表性工作[49]。在无源域、极端小样本条件下，如何在保证生成图像逼真度的同时大幅提升多样性，实现两者的良好平衡，依然是该领域的核心难点。

1.4.4 图像风格迁移研究现状

图像风格迁移（Image Style Transfer）作为一种特殊的图像生成任务，致力于将一种图像的风格特征迁移至另一幅图像的内容上，生成具有特定艺术风格或语义信息的图像。在风格迁移中，“内容”是指源图像中的各种实例，包括其语义信息、轮廓结构和色彩特征。长时间以来，风格迁移主要依赖于提供风格参考图像的图像引导方法，能够将源图像转变为具备经典艺术作品风格的图像。在深度学习发展的推动下，图像引导的风格迁移经历了从传统纹理建模[50]到神经风格迁移（Neural Style Transfer，NST）的转变[51-52]。GAN 因其强大的图像生成能力被迅速用于风格迁移任务。例如 CycleGAN 利用循环一致性损失与对抗损失实现无配对风格迁移[51,53]。CUT 方法则进一步通过对比学习保持输入输出图像之间结构一致性并最大化局部区域的互信息[52]。Zhao 等[54]则提出在 GAN 中引入多注意力机制，专注于照片卡通化。采用 CNN 或 Transformer 的编码器-解码器结构也为风格迁移提供了新的技术路径。AdaIN 使用 VGG 网络编码风格与内容特征，通过调整特征空间中的均值与方差来对齐风格[55]。AdaAttN[56]与 CovAttn[57]则动态地根据内容与风格特征调整注意力权重，实现更精准的风格迁移。CAST 方法利用对比学习分析不同风格之间的相似性与差异性，从而提升风格迁移的精准度[58]。基于 ViT[59]的方法，如 StyTr2[60]也展现出生成风格化内容的能力。近年来，DM 在各项图像相关领域均表现突出，包括风格迁移领域，DDIB 通过独立训练内容域与风格域的 DM，并在潜在空间桥接两者，生成风格化图像[61]。StyleDiffusion[62]与 InST[63]也提出分别处理内容提取与风格融合的双扩散模型架构，以更好地实现风格解耦与迁移。此外，OSASIS 提出结构保持网络以融合域外风格参考图像[64]。然而，图像引导的方法受限于特定的参考图像，限制了其更广泛的应用，并且多数方法还需要基于大规模风格图像集[65]的额外模型训练或微调，过程较为复杂且效率较低。此外，寻找适合的风格参考图像也较为耗时，尤其是在需要迁移某些无法精确匹配图像风格的场景下。

近年来，随着大型视觉-语言模型，如 CLIP[11,66]的发展，推动了零样本文本引导风格迁移（Zero-shot Text-guided Style Transfer）的研究。其目标是让模型仅凭文本描述，就能将图像风格化为训练时未明确见过的新颖风格，其核心在于模型的零样本泛化能力，这与需要少量样本进行适配的小样本风格注入技术，如 Dreambooth[67]，LoRA[68]等有本质区别。主流方法通常利用 CLIP 等模型的引导信号来指导预训练

的生成模型，可以选择预训练的 StyleGAN[12,69]、VQGAN[70] 或扩散模型[71-72]。其中，基于扩散模型的方法通过对比损失引导如 ZeCon[73]，或免训练的过程调制，如 FreeStyle[74] 等技术来实现文本到风格的转换。然而，零样本文本引导风格迁移仍面临严峻挑战：(1) 内容与风格的平衡：如何在强风格注入下最大限度地保留原始图像内容结构；(2) 文本指令的忠实度：模型对复杂、抽象或组合式文本描述的理解与视觉转化能力有限，可能导致风格偏差。

近来基于注意力操控[75-76] 的图像编辑领域的发展也显示出在风格迁移方面的潜力。Prompt Tuning Inversion (PTI) 通过提示词调整和无分类器引导（Classifier-free Guidance）实现了精确的文本驱动编辑[77]。InfEdit 利用特殊的方差调度和统一的注意力控制机制，实现了无需反演的编辑[78]。PnP Inversion 分离了源扩散分支和目标扩散分支，通过最小化的修改显著提升了编辑性能和计算效率[79]。然而，由于这些模型通常面向更广泛的图像编辑任务，针对风格迁移这一特定目标可能不如专门设计的模型有效或控制精确。

1.5 论文主要工作

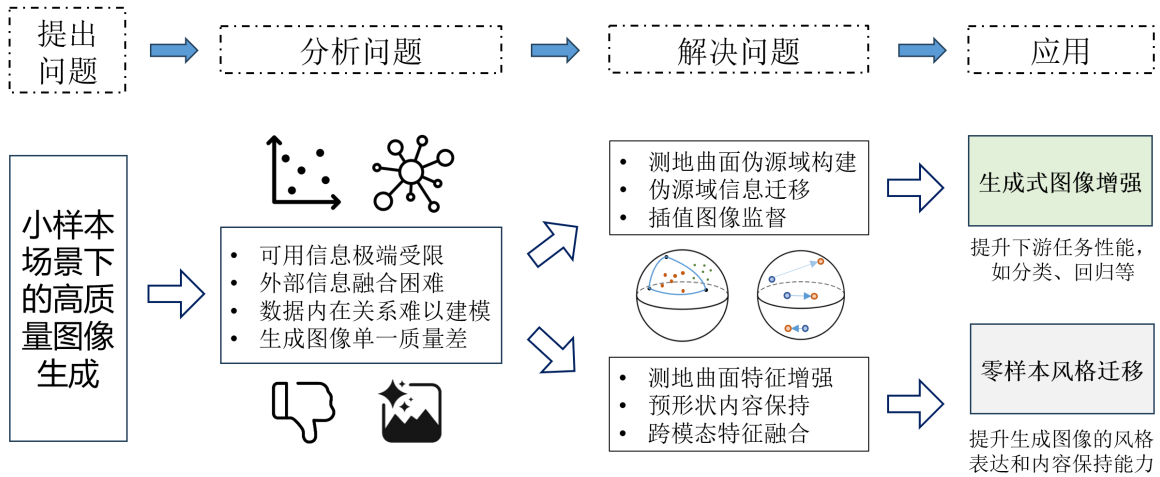


图 1.1 本论文的技术路线示意图

为探索在小样本环境下实现高质量图像生成的有效途径，本论文基于形状空间理论与计算机视觉技术，提出并研究了以下两个核心方法，技术路线如图 1.1 所示。主要研究工作与创新点概括如下：

（1）提出一种基于预形状空间测地曲面信息迁移的小样本图像生成方法，用于

克服现有小样本生成方法过度依赖大规模预训练或在极端稀疏数据下多样性与保真度难以平衡的问题。该方法在无需外部大规模数据集或预训练模型的前提下，通过测地曲面特征增强模块，仅利用目标任务自身的少量训练样本在预形状空间中构建测地曲面，并在此曲面上进行非线性特征增强，以模拟更丰富的数据分布，实现信息迁移。结合插值监督与正则化约束，该方法旨在无需外部数据即可显著提升生成图像的质量和多样性，并缓解模式坍塌。实验证明，该方法在多个数据集上表现优越，且生成的样本证实可有效提升下游预测任务性能。

(2) 提出一种基于形状空间特征增强的零样本场景下文本引导风格迁移方法，应对现有文本引导风格迁移方法在内容保持与风格控制方面的挑战，旨在改善风格控制与内容保持的平衡。该方法将预形状空间测地曲面特征增强思想应用于扩散模型，以促进文本风格与图像内容特征的有效融合。同时，引入局部信息处理并设计了预形状自相关一致性约束来加强内容结构的稳定性。实验证明，该方法能在无需额外训练或风格参考下，实现灵活的文本引导风格控制，同时较好地保持原始内容结构。

1.6 论文组织结构

针对小样本图像生成与风格迁移中的数据稀缺问题，本论文提出了一种基于形状空间理论的特征增强方法，旨在提升小样本条件下图像生成的质量与多样性，并验证其在材料图像生成与风格迁移等实际应用中的有效性。论文结构安排如下:

第二章介绍了本论文的相关理论与技术基础，主要包括图像生成模型的基本理论、生成对抗网络与扩散模型的相关原理、以及形状空间理论的基本概念和应用。还包括文本-图像多模态预训练模型和与生成图像质量相关的评价指标。

第三章聚焦于基于预形状空间中测地曲面增强的小样本图像生成方法。首先，本章概述了该方法的整体思路，依次介绍了基于测地曲面的伪源域构建、伪源域到目标域的信息迁移、插值监督与正则化模块以及最终优化目标函数的设计。随后，通过实验分析部分展示了实验设置、定性对比、定量对比、消融实验以及生成图像有效性分析，全面验证了方法在小样本条件下提升图像生成质量与多样性的效果。

第四章提出了基于预形状空间中测地曲面增强的零样本文本引导图像风格迁移方法。该章首先阐述了基于文本信息引导的扩散模型风格迁移的基本思路，接着介绍了基于测地曲面特征增强模块特征融合的风格控制及在预形状空间上实现内容保

持的策略。实验分析部分包括实验设置、针对想象类风格与常见类风格的风格迁移模型和图像编辑模型对比实验，以及对于提出方法各个组成项的消融实验，验证了该方法在保持图像内容一致性与风格化一致性方面的优势。

第五章总结了全文的工作，并对未来的研究方向进行了展望。分析了本论文提出的方法在小样本图像生成与零样本文本引导风格迁移领域的优势与不足，并对未来可能的研究方向做了展望。

第二章　相关理论和技术概述

本章旨在为后续章节中提出的、基于形状空间理论的小样本生成与风格迁移方法提供必要的背景知识。为此，本章将依次介绍形状空间理论基础、相关的深度生成模型、多模态技术以及主要的性能评价指标，从而为理解后续方法创新与实验验证奠定所需的技术基础。

2.1　形状空间理论

形状空间理论由 Kendall 于 1984 年在几何数据分析领域提出[19]。该理论将几何形状定义为移除了平移、缩放和旋转等刚性变换影响后所保留的本质结构信息，对后续相关研究产生了重要影响。

在二维欧式空间中，一个形状 P 可以通过一组标志点（Landmarks）表示，形式为 $P = \{p_1(x_1, y_1), \ldots, p_m(x_m, y_m)\} \in \mathbb{R}^{2\times m}$。在将形状 P 投影到预形状空间时，需先进行均值消减（Mean Reduction）操作 $\mathcal{Q}(\cdot)$ 以去除坐标位置对于形状结构信息的影响：

$$P' = \mathcal{Q}(P) = \{p'_i = (x_i - \bar{x}, y_i - \bar{y})\}, \tag{2.1}$$

其中 $i = 1, \ldots, m$ 且 m 表示标志点的数量，$\bar{x}$ 和 $\bar{y}$ 分别对应 $\{x_i\}$ 和 $\{y_i\}$ 的均值。接下来通过归一化（Normalization）操作 $\mathcal{V}(\cdot)$，去除缩放影响，从而得到预形状 τ：

$$\tau = \mathcal{V}(\mathcal{Q}(P)) = \mathcal{V}(P') = \frac{P'}{\|P'\|}, \tag{2.2}$$

其中 $i = 1, \ldots, m$ 且 m 表示标志点的数量，$\|\cdot\|$ 表示欧几里得范数。经过上述投影操作，所有以 m 个标志点表示的预形状都被嵌入到一个单位超球面 S_*^{2m-3} 上，即预形状空间。任何的预形状都是超球面上的一个点或一个向量，并且一个形状所有的旋转变换都在该超球面的一个大圆上，记为 $O(\tau)$，作为形状空间中的一个点或向量。形状空间则预形状空间中所有大圆组成的集合定义：

$$\Sigma_2^m = \{O(\tau) : \tau \in S_*^{2m-3}\}, \tag{2.3}$$

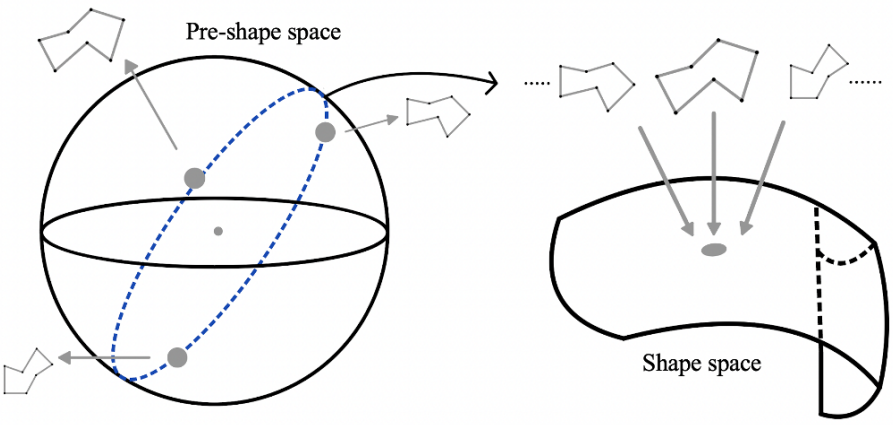


图 2.1　预形状空间和形状空间的示意图[26]

预形状空间和形状空间的示意图如图 2.1 所示。要将 P 投影到形状空间，需要在复数域中执行较为复杂的运算，因此大多数研究主要聚焦于预形状空间。欧式空间中使用欧式距离计算两点之间的最短距离，对于预形状空间 S_*^{2m-3}，使用测地距离衡量两个预形状之间的最短距离：

$$d\big(\tau_1, \tau_2\big) = \arccos(\langle \tau_1, \tau_2 \rangle), \tag{2.4}$$

其中 $\tau_1, \tau_2 \in S_*^{2m-3}$ 是两个预形状，$\langle \cdot, \cdot \rangle$ 表示点乘。

在预形状空间中，可以用一条测地曲线（Geodesic Curve）连接两个预形状，这条测地曲线是超球面上的大圆弧，代表了两点间的最短路径。测地曲线上的点可用于描述一组变化的形状或生成中间形状。Han 等人[21-22] 提出了在预形状空间中，基于测地曲线生成两个之间更多新的预形状的方法。连接 τ_1 和 τ_2 的测地曲线可通过球面线性插值（Spherical Linear Interpolation，SLERP）定义：

$$\mathbb{G}_{cur}\left(\tau_1, \tau_2\right)(s) = \left(cos(s)\right)\tau_1 + \left(sin(s)\right)\frac{\tau_2 - \tau_1 cos(d(\tau_1, \tau_2))}{sin(d(\tau_1, \tau_2))}, \tag{2.5}$$

其中弧度 s（$0 \leq s \leq d(\tau_1, \tau_2)$）可控制新生成数据点与 τ_1 的测地距离。通过逐步增大 s，可以生成一系列由 τ_1 逐渐变化到 τ_2 的预形状。

如果只有两个预形状，就可以直接基于公式 (2.5) 生成新的预形状。然而，当需要表示或拟合超过 2 个预形状的变化或分布时，仅依赖公式 (2.5) 所示的连接任意两点的测地曲线就显得不足，因为它无法捕捉一组形状的整体变异性或中心趋势[80]。

一种方法是寻找一条最优测地曲线，它能最好地代表这组数据的中心路径或主

要变化方向。这个概念被形式化为主测地分析（Principal Geodesic Analysis，PGA）[80]，它是欧氏空间主成分分析（Principal Component Analysis，PCA）到如预形状空间等黎曼流形上的推广。PGA 旨在找到一条测地曲线，使得所有输入点到该结构的平方测地距离之和最小。实践中，计算精确的 PGA 可能很复杂。因此，一种常见的近似方法是切空间主成分分析（Tangent PCA，tPCA）[80]。该方法通过将数据映射到局部切空间执行 PCA 再映射回流形来近似主测地曲线，但作为一种基于切空间近似的方法，其精度会受到流形几何特性的影响。为避免这种近似可能带来的问题，FAGC 提出了直接在预形状空间 S_*^{2m-3} 上寻找最优测地曲线的方法[26]。其直接优化定义测地曲线在预形状空间中的两个端点。该方法利用流形的内在结构，例如球面插值公式 (2.5)，以及为特定任务设计的损失函数，从而避免了 tPCA 中显式的切空间映射步骤及其潜在的近似失真。

当需要捕捉数据中更复杂的多维变化模式时，单一测地曲线可能不足，因而引出了构建更高维度的测地曲面或子流形的需求。从概念上讲，这些测地结构，即测地曲线与测地曲面，可以被视为预形状空间这种弯曲流形上对欧氏空间中直线和平面的自然推广。虽然一种常见方法也是基于切空间近似来定义这些高维曲面[24,80]，但这种方法存在固有的局限性。由于预形状空间 S_*^{2m-3} 具有正曲率[81]，切空间仅为局部线性近似，导致该方法依赖的对数和指数映射会引入失真，尤其当数据分布广泛或远离均值点时，其保真度有限[82]。为了克服这种基于近似方法的局限性，Pennec[83] 提出了 Fréchet 重心子空间（Fréchet Barycentric Subspaces，FBS）这一概念。FBS 完全基于流形上的点和内在测地距离进行定义，不涉及切空间映射。一个 FBS 被定义为由 n 个参考点 $\{\tau_1, \ldots, \tau_n\}$ 通过加权 Fréchet 均值生成的所有点的集合 μ，如公式 (2.6) 所示：

$$\mathbb{G}_{FBS}(\tau, \omega) = \left\{ \arg\min_{\mu} \sum_{j=1}^{n} \omega_j d(\mu, \tau_j) \;\middle|\; \sum_{j=1}^{n} \omega_j \neq 0 \right\}, \tag{2.6}$$

其中 $\omega = \{\omega_1, ..., \omega_n\}$ 是满足特定条件的权重集。通过改变权重，可以在由参考点定义的内在、通常非线性的子空间内移动。FBS 的核心优势在于其直接在预形状空间中操作，避免了切空间投影相关的近似误差与失真，因此有望更准确地捕捉数据的真实内在结构，特别是在处理分布复杂或流形曲率影响显著的数据时[83]。

无论是沿着最优测地曲线，还是在通过 FBS 定义的测地曲面上进行采样，都可以在预形状空间中生成一系列新的且有效的预形状。这些生成的预形状可以看作是

原始给定预形状之间基于测地距离的等价变换。考虑到从图像数据中提取的特征通常也包含一定的内在结构信息，将这些特征投影至预形状空间，并利用测地结构生成新的预形状特征，可以作为一种有效的数据集扩充手段。因此，利用形状空间理论，特别是其测地结构，生成新数据是解决图像小样本学习问题的一个有潜力的途径。本论文旨在探索和设计在预形状空间中利用这些方法生成新数据点，以拟合数据集中的图像特征分布，实现数据增强，从而更有效地挖掘出小样本数据集中的信息。

2.2 图像生成模型

图像生成（Image Generation）是一项重要的计算机视觉任务，旨在利用深度学习模型从随机噪声或潜在空间中生成逼真的图像。目前，主流的图像生成模型主要包括变分自编码器（Variational Autoencoder，VAE）、生成对抗网络（Generative Adversarial Network，GAN）和扩散模型（Diffusion Model，DM）。这些生成模型不仅用于图像合成，也广泛应用于图像风格迁移、图像修复、超分辨率、数据增强等下游任务。本论文的研究工作主要基于生成对抗网络和扩散模型展开，因此本节将重点介绍这两类模型的相关理论与技术。

2.2.1 生成对抗网络

生成对抗网络（GAN）[84] 是深度生成模型的重要分支，其核心思想源于零和博弈。一个标准的 GAN 包含两个相互竞争的神经网络：生成器 $\mathcal{G}$ 和判别器 $\mathcal{D}$。生成器 $\mathcal{G}$ 负责学习真实数据的分布，它通常接收一个从先验分布 $p(z)$ 中采样的随机噪声向量 z 作为输入，并尝试生成尽可能逼近真实数据分布 $\mathbb{D}_{real}$ 的样本 $\mathcal{G}(z)$。判别器 $\mathcal{D}$ 则作为一个二元分类器，负责判断输入样本是来自真实数据集 $\mathbb{D}_{real}$ 还是由生成器 $\mathcal{G}$ 产生的，其输出 $\mathcal{D}(x)$ 表示样本 x 为真实的概率。$\mathcal{G}$ 和 $\mathcal{D}$ 在训练过程中进行对抗：$\mathcal{G}$ 的目标是生成能欺骗 $\mathcal{D}$ 的图像，而 $\mathcal{D}$ 的目标是准确区分真实图像 x 和生成图像 $\mathcal{G}(z)$。这种对抗过程可以通过优化以下两个损失函数来交替进行：

$$L_{adv}^{\mathcal{G}} = -\mathbb{E}_{z\sim p(z)}[\log(\mathcal{D}(\mathcal{G}(z)))], \tag{2.7}$$

以及

$$L_{adv}^{\mathcal{D}} = \mathbb{E}_{x\sim\mathbb{D}_{real}}[\log(1-\mathcal{D}(x))] + \mathbb{E}_{z\sim p(z)}[\log(\mathcal{D}(\mathcal{G}(z)))]. \tag{2.8}$$

尽管基础 GAN 框架具有开创性意义，但其原始形式的训练过程常面临诸如梯度消失、模式坍塌（Mode Collapse）以及训练不稳定等挑战，这些问题在处理复杂高维数据或样本有限时尤为突出。为克服这些难题，研究者们提出了一系列关键的改进模型。例如，WGAN（Wasserstein GAN）[85] 及其改进 WGAN-GP[86] 通过引入 Wasserstein 距离度量和梯度惩罚，显著提升了训练的稳定性和收敛性，并缓解了模式坍塌问题。渐进式增长 GAN（Progressive GAN，ProGAN）则采用由低到高、逐步增加网络层数和训练图像分辨率的策略，生成了高分辨率的逼真图像[87]。BigGAN 通过采用更大的模型容量、增大批处理大小、引入截断技巧、谱归一化和自注意力机制等手段，在大规模数据集上生成了具有更高保真度和多样性的图像[88]。

在众多 GAN 变体中，StyleGAN 系列[89-92] 因其在生成高分辨率、高质量图像以及提供优越的风格控制和编辑能力方面取得的巨大成功，已成为该领域最具影响力的代表性工作之一。鉴于本论文后续章节的研究工作采用了 StyleGAN2 作为基础架构之一，在此对其核心设计理念进行简要介绍，StyleGAN2 的网络结构示意图如图 2.2 所示。

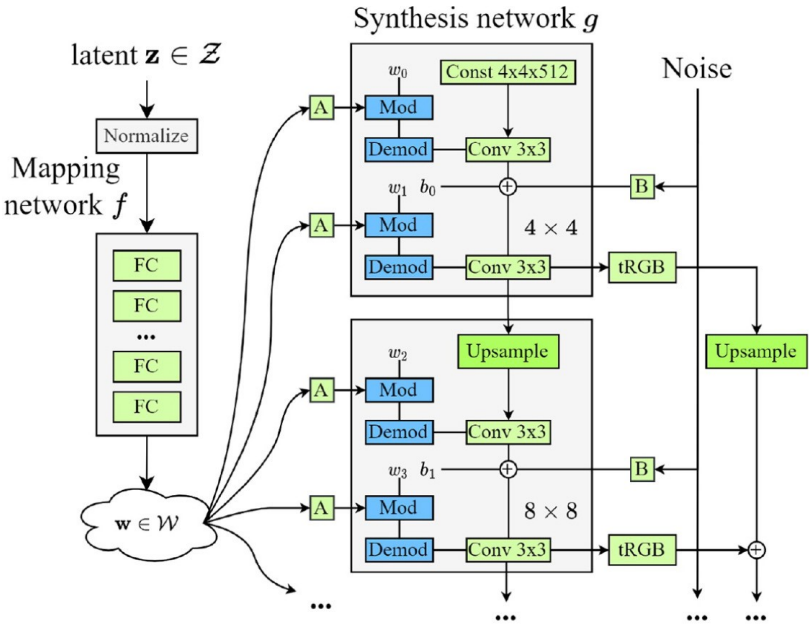


图 2.2 StyleGAN2 的网络结构示意图[90,93]

StyleGAN2[90] 的生成器由两部分组成：映射网络（Mapping Network）和合成网络（Synthesis Network）。映射网络是一个由多个全连接层构成的深度网络，通常为 8 层 MLP，它将输入的标准正态分布潜变量 $z \in \mathcal{Z}$ 映射到一个解耦性更好的中间潜变量 $w \in \mathcal{W}$，目的是让 w 的不同部分能更独立地控制生成图像的不同层次属性。合成网络则负责从 w 生成最终图像。它不直接使用 w 作为输入，而是从一个可学习的 $4 \times 4 \times 512$ 常量张量开始，每个分辨率级别包含两个卷积层、上采样、加入噪声、激活函数等，通过一系列卷积模块逐步提升特征图的空间分辨率直至目标尺寸。其结构上的关键创新在于：(1) 风格注入方式：采用权重调制与解调（Weight Modulation and Demodulation）技术。具体来说，卷积层的权重会根据从 w 导出的风格向量进行逐通道缩放（调制），然后进行归一化（解调）以保持特征统计量的稳定，这种方式取代了原始 StyleGAN 中的 AdaIN，有效提升了图像质量并消除了特定伪影。(2) 随机细节引入：在每个卷积层之后，向特征图中加入学习到的随机噪声输入，这些噪声独立作用于每个像素，用于生成非结构化的随机细节。(3) 多分辨率融合：通过跳跃连接（Skip Connections）将每个分辨率级别输出的 RGB 特征图累加到最终的图像上，有效融合多尺度信息。

StyleGAN2 的判别器采用了残差网络（Residual Network，ResNet）[94] 架构，由一系列包含下采样操作的卷积块和残差连接构成，并同样使用跳跃连接来整合不同分辨率的特征，以有效地区分真实图像和生成图像。除了上述网络结构的改进，StyleGAN2 还引入了重要的设计变更与正则化策略。它显式地移除了渐进式增长训练机制，并通过优化的网络结构和训练策略来保证高分辨率训练的稳定性。同时，它引入了路径长度正则化（Path Length Regularization，PLR），通过惩罚 $\mathcal{W}$ 空间中扰动引起的图像空间剧烈变化，来鼓励潜空间到图像空间的映射更加平滑和解耦，这不仅提升了生成图像的感知质量和一致性，也改善了潜空间的可编辑性。这些精心设计的结构和策略共同构成了 StyleGAN2 强大的图像生成能力。

除了上述主流改进方向，研究者也在探索将其他理论或结构融入 GAN 框架。例如，利用最优传输理论改进判别器设计以稳定训练[95]，以及结合 VAE 的优点，形成如 CVAE-GAN[96]、HP-VAE-GAN[37] 等混合模型，试图兼顾生成质量与模式覆盖度。近年来，Transformer 架构[97] 的成功也启发了新的 GAN 设计思路，如完全基于 Transformer 的 TransGAN[98]，以及结合 CNN 进行图像分词，再用 Transformer 建模全局依赖关系的 VQGAN[99]，这些工作展现了不同架构在图像生成任务中的潜力与

特性。GAN 及其变体的发展，为高质量图像生成提供了强大的工具集，但也持续面临训练稳定性、模式覆盖度和可控性等方面的挑战，尤其是在小样本或特定控制任务中。

2.2.2 扩散模型

去噪扩散概率模型（Denoising Diffusion Probabilistic Model，DDPM）[2-3] 是近年来在生成模型领域取得显著成功的一类方法。与通过生成器和判别器进行对抗博弈的 GAN 不同，扩散模型通过精心设计的前向扩散（Forward Diffusion）过程和反向去噪（Reverse Denoising）过程，在理论上建立了从复杂数据分布到简单先验分布，通常是标准高斯分布，之间的可逆转换，从而实现了稳定且高质量的图像生成。

具体来说，DDPM 首先定义了一个前向扩散过程，该过程逐步将一个来自真实数据分布 $q(x_0)$ 的样本 x_0 ，在 T 个离散时间步内，通过反复添加少量高斯噪声，最终转化为一个近似服从标准高斯分布 $\mathcal{N}(0, I)$ 的噪声潜变量 x_T。这个过程通常通过一个预设的方差调度（Variance Schedule）$\beta_t \in (0, 1)$，其中 $t = 1, \ldots, T$，来控制每一步添加噪声的幅度，其单步转移概率为：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{2.9}$$

扩散过程的一个重要性质是，可以利用重参数化技巧，直接从初始样本 x_0 采样得到任意时刻 t 的含噪样本 x_t，其形式为：

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{2.10}$$

其中 $\epsilon \sim \mathcal{N}(0, I)$ 是标准高斯噪声，$\alpha_t = 1 - \beta_t$，且 $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ 是累积乘积系数。这个公式清晰地显示了 x_t 是原始信号 x_0 和噪声 ϵ 的线性组合，其权重由累积系数 $\bar{\alpha}_t$ 决定。

生成图像的过程对应于扩散模型的反向过程，即从 $x_T \sim \mathcal{N}(0, I)$ 出发，通过神经网络学习并逐步去除噪声，最终恢复出符合真实数据分布的样本 x_0。这个反向过程的目标是近似真实的后验概率 $q(x_{t-1}|x_t, x_0)$，通常由一个参数为 θ 的神经网络，常用 ϵ_θ 表示，来建模条件概率 $p_\theta(x_{t-1}|x_t)$。在 DDPM 框架下，反向去噪的单步操作可

以表示为：

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z, \tag{2.11}$$

其中，$z \sim \mathcal{N}(0, I)$ 是标准高斯噪声，$\epsilon_\theta(x_t\ t)$ 代表神经网络在给定含噪图像 x_t 和时间步 t 的条件下，对添加到 x_0 上的原始噪声 ϵ 的预测。σ_t 控制了采样过程中的随机性。通过从 $t = T$ 到 $t = 1$ 迭代执行这个去噪步骤，模型就能从纯噪声生成一张清晰的图像。

扩散模型的训练目标是优化神经网络 ϵ_θ，使其能准确预测添加到 x_0 上的噪声。这通常通过最小化一个简化的损失函数来实现，该函数对应于最大化证据下界的一个变分项：

$$\min_\theta \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(1,T)} \left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2 \right], \tag{2.12}$$

该损失函数直观地表示为，在所有时间步 t 上，让网络预测的噪声 ϵ_θ 与实际添加的噪声 ϵ 之间的均方误差最小。

近两年来，扩散模型在多项图像生成基准测试中展现出与 GAN 相当甚至更优的性能，尤其在生成图像的多样性和训练过程的稳定性方面具有明显优势，使其在许多场景，如无条件生成、文本到图像生成等，成为主流方法之一。然而，扩散模型的主要缺点在于其采样过程通常需要较多的迭代步数，导致生成图像的速度远慢于单步生成为主的 GAN。此外，训练高质量的扩散模型通常需要大量的计算资源和时间。为了提高扩散模型的效率和性能，研究者提出了多种改进方法。例如，基于分数的生成模型（Score-based Generative Models，SGMs）[100] 从随机微分方程的角度统一了扩散过程和分数匹配（Score Matching）方法，通过估计数据在不同噪声尺度下的对数概率密度梯度，也称为分数，为理解和设计生成模型提供了新的视角。

除了无条件图像生成，扩散模型在条件生成任务，特别是文本引导的图像生成（Text-to-image Synthesis）领域取得了突破性进展。诸如 DALL · E 2[101] 和 Stable Diffusion[5] 等模型，能够依据用户提供的文本描述，生成与之语义高度一致且视觉质量极高的图像。这些模型通常将文本提示（Text Prompt）编码为条件信息，并通过交叉注意力等机制将其有效地融入到 ϵ_θ 网络的去噪过程中，从而实现对生成内容的精确控制。近期的研究进一步探索了更精细的控制方式来指导生成过程，进一步提升了扩散模型在可控图像合成方面的能力[102]。

鉴于本论文第四章的研究工作基于无条件扩散模型展开，其核心的去噪网络 ϵ_θ 通常采用 U-Net 架构[103] 或其变体[3,104]。U-Net 因其对称的编码器-解码器结构和跳跃连接而特别适用于图像生成任务，能够有效结合多尺度信息，同时捕捉局部细节和全局上下文。为了使去噪过程依赖于当前的时间步 t，模型引入了时间步嵌入（Timestep Embedding）机制，通常将 t 通过过正弦位置编码映射为一个向量，并通过小型 MLP 处理后，以尺度和偏置的形式调制网络内部的残差卷积块。此外，为了建模图像内部的长距离依赖关系以确保全局结构的一致性，U-Net 的较低分辨率层级通常会集成多头自注意力（Multi-Head Self-Attention）模块。这些关键组件，包括 U-Net 骨架、时间步条件化、残差学习和自注意力机制协同工作，使得扩散模型能够有效地学习从噪声到数据的复杂映射，生成高质量、多样化的图像。

2.3 文本-图像多模态预训练模型

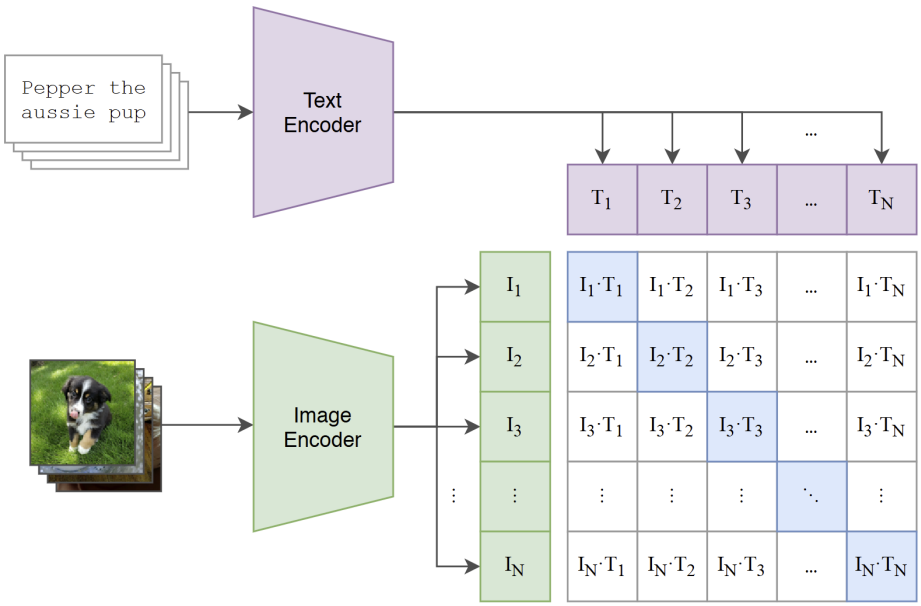


图 2.3 CLIP 的训练方式[11]

随着多模态学习（Multimodal Learning）的兴起，对图像和文本这两种核心信息模态进行联合处理已成为研究热点之一。这类方法通常需要构建两个独立的编码器，分别用于编码图像和文本信息，并在一个共享的空间中对齐它们的表示。其中，由 OpenAI 于 2021 年提出的 CLIP（Contrastive Language-Image Pre-Training）[11] 模型，是基于对比学习（Contrastive Learning）范式的一个里程碑式工作。该模型通过在海量

图文数据上进行预训练，学习到了强大的图像-文本联合表示能力，在多种下游视觉任务中展现出卓越的零样本泛化性能，其训练方式如图 2.3 所示。

与传统的监督学习模型依赖大量带有特定类别标签的图像数据不同，CLIP 的核心思想是通过从互联网收集的海量原始图文对数据进行预训练。它利用这些自然存在的对应关系，学习图像和文本之间的潜在语义联系。CLIP 模型包含一个图像编码器（Image Encoder）和一个文本编码器（Text Encoder)。图像编码器可以选用经典的卷积神经网络，如 ResNet 或 ViT 系列，文本编码器则通常采用标准的 Transformer 架构。

CLIP 训练的核心目标是学习一个共享的多模态嵌入空间，其优化过程主要由对比损失函数驱动。具体而言，对于一个批次内的 N 个图文对样本，模型首先使用图像编码器和文本编码器分别提取它们的特征向量。然后，计算所有图像特征与所有文本特征之间的余弦相似度，形成一个 $N \times N$ 的相似度矩阵。在这个矩阵中，对角线元素代表了正确匹配的图文对（即正样本对），而非对角线元素则代表了不匹配的图文对（即负样本对）。CLIP 的优化目标是通过对比损失函数，最大化正样本对之间的相似度，同时最小化负样本对之间的相似度。通过这种方式，模型被驱动去学习一个共享的多模态嵌入空间，在该空间中，语义相关的图像和文本表示彼此靠近，而语义无关的则相互远离。

得益于在大规模、多样化的网络数据上的预训练，CLIP 模型学习到了强大的、具有良好泛化能力的视觉和语言联合表示。这使得 CLIP 在各种下游任务中表现出卓越的零样本迁移能力。例如，在图像分类任务中，研究者无需为特定数据集进行模型微调。只需为每个类别构造相应的文本描述，然后计算待分类图像的特征与所有类别文本描述特征之间的相似度，并将图像归类到相似度最高的那个类别即可。这种零样本分类能力极大地降低了对标注数据的依赖。此外，CLIP 的预训练表示也被证明在图像-文本检索、作为图像生成模型的引导信号、视觉问答等多种跨模态任务中非常有效，展现了强大的应用潜力。

因此，CLIP 的提出不仅显著推动了多模态预训练技术的发展，减少了对特定任务标注数据的需求，也为后续的视觉-语言研究提供了重要的基准模型和强大的基础表示。

2.4 评价指标

为了定量评估本论文所提出方法的性能，实验部分主要采用了弗雷歇初始距离（Fréchet Inception Distance，FID）[105]、学习感知图像块相似度（Learned Perceptual Image Patch Similarity，LPIPS）[106]、峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）、结构相似性指数（Structural Similarity Index Measurement，SSIM）[107] 以及 CLIP 分数（CLIP-score）[11,108] 等指标。其中，FID 与 LPIPS 主要用于评估生成图像的整体质量和感知相似性，而 PSNR、SSIM 与 CLIP-score 则侧重于评估风格迁移任务中生成图像在内容保持、结构相似性以及风格语义符合度方面的表现。LPIPS 在风格迁移任务中也可用于衡量内容保持能力。

FID 是评估生成模型生成图像真实性和多样性的常用指标。它通过一个预训练的 Inception V3 网络提取真实图像集和生成图像集的高维特征，并假设这些特征向量服从多元高斯分布。然后计算这两个高斯分布之间的弗雷歇距离（Fréchet Distance），距离越小表示两个分布越接近。其计算公式为：

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{2.13}$$

其中，μ_r 和 Σ_r 分别是真实图像特征集的均值向量和协方差矩阵，μ_g 和 Σ_g 则是生成图像特征集的均值向量和协方差矩阵。较低的 FID 值通常意味着生成图像的分布与真实图像分布更相似，即生成质量更高、多样性更好。

LPIPS 旨在衡量两张图像之间的感知相似度，使其更符合人类的视觉感受，而不是仅仅依赖像素级别的差异。该指标利用预训练的深度卷积网络，如 VGG 或 AlexNet，提取图像在不同网络层级的特征图。然后，计算两张图像在对应层级特征图上的差异，并通过学习到的权重 w_l 对不同层级、不同通道的差异进行加权求和：

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{x}_l^{hw} - \hat{y}_l^{hw})\|_2^2, \tag{2.14}$$

式中，x 和 y 分别代表待比较的两张图像，$\hat{x}_l^{hw}$ 和 $\hat{y}_l^{hw}$ 是在第 l 层位置 (h, w) 处提取并归一化的特征向量，w_l 是该层学习到的通道权重，$\odot$ 表示逐元素相乘。LPIPS 值越小，表明两张图像在感知上越相似。在风格迁移任务中，LPIPS 也常被用来评估内容保持的效果，此时通常计算风格化后的图像与原始内容图像之间的 LPIPS 距离。较低的 LPIPS 值表明迁移后的图像在引入新风格的同时，在感知层面上更好地保留

了原始内容图像的结构与特征。

PSNR 是一种广泛使用的图像质量评价指标，它基于图像间的均方误差（Mean Squared Error，MSE）计算。PSNR 衡量的是信号的最大可能功率与破坏信号精度的噪声功率之间的比率，单位为分贝（dB)。其定义为：

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{2.15}$$

其中，MAX_I 是图像像素值的最大可能范围，对于 8 位灰度或彩色图像通常是 255。MSE 的计算方式为：

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2, \tag{2.16}$$

这里 I 和 K 分别代表原始内容图像和风格迁移后的图像，尺寸为 $m \times n$。PSNR 值越高，表示迁移后的图像与原始内容图像在像素级别上的失真越小，内容保持得越好。

SSIM 则从图像结构信息的角度出发来衡量两幅图像的相似性。相比于 PSNR 只关注像素误差，SSIM 认为人类视觉系统更关注图像中的结构信息、亮度和对比度。它综合了亮度比较 $l(x,y)$、对比度比较 $c(x,y)$ 和结构比较 $s(x,y)$ 三个方面：

$$\text{SSIM}(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}, \tag{2.17}$$

通常简化形式，即设设 $\alpha = \beta = \gamma = 1$，并使用局部统计量计算为：

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{2.18}$$

其中，μ_x, μ_y 分别是图像 x, y 的局部均值，σ_x^2, σ_y^2 是局部方差，σ_{xy} 是局部协方差。c_1, c_2 是用于维持稳定性的常数。SSIM 的取值范围通常在 $[-1, 1]$ 之间，当用于衡量与原始图像的相似度时，值越接近 1，表示两幅图像的结构越相似，视觉效果越接近。

CLIP-score 是近年来利用强大的 CLIP 模型来评估图像与文本描述之间语义一致性的指标。在风格迁移任务中，它可以用来衡量迁移后的图像 I 在多大程度上符合目标风格的文本描述 t。计算方法是利用预训练的 CLIP 模型分别提取图像 I 的特征 $E_I(I)$ 和文本 t 的特征 $E_T(t)$，然后计算这两个特征向量之间的余弦相似度：

$$\text{CLIP-score}(I,t) = \frac{\langle E_I(I), E_T(t) \rangle}{\|E_I(I)\| \cdot \|E_T(t)\|}, \tag{2.19}$$

CLIP-score 值越接近 1，意味着风格迁移后的图像在语义层面上与目标风格描述更加匹配，体现了更好的视觉-文本一致性或风格符合度。

通过综合运用上述这些评价指标，本论文能够从生成图像的真实性（FID)、多样性（LPIPS)、内容保真度（PSNR，LPIPS)、结构相似性（SSIM）以及风格语义符合度（CLIP-score）等多个维度，对所提出的图像生成与风格迁移方法进行更全面、更客观的性能评估。

2.5 本章小结

本章阐述了本论文使用到的关键技术和基础理论。首先介绍形状空间理论。其次，介绍图像生成模型，包括生成对抗网络和扩散模型以及一些经典的网络结构。接着，是对文本-图像多模态预训练模型的介绍。最后，为了客观地评价本论文的方法，本章介绍了使用的评价指标。

第三章　基于预形状空间中测地曲面信息迁移的小样本图像生成

本章聚焦于在无大规模预训练支持的极端小样本图像生成场景。针对该场景下生成图像质量与多样性不足的瓶颈，本章提出了一种基于小样本场景的数据增强方法，通过引入预形状空间上的测地曲面，通过特定的插值与特征对齐技术，突破小样本图像生成中的瓶颈。该方法可在多样化的小样本数据集上进行训练，生成具有高保真度和多样性的图像。通过利用生成的图像进行数据集扩充，本方法能够有效应对特定专业领域中的小样本学习任务，进一步提升模型的性能与泛化能力。

3.1　方法概述

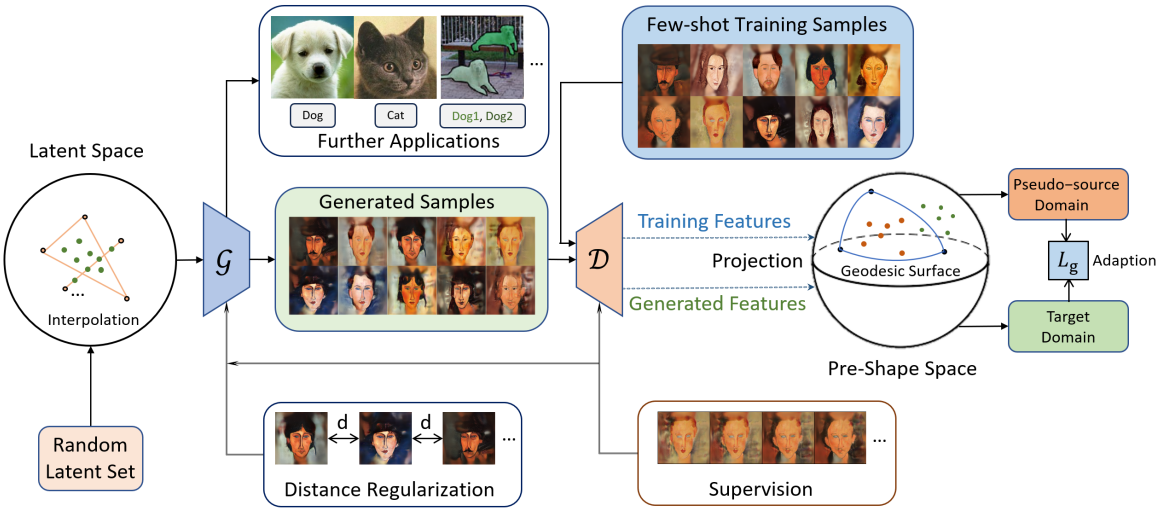


图 3.1　基于预形状空间中测地曲面信息迁移的小样本图像生成方法的流程图。该方法利用极少量训练样本，通过流形数据增强构建一个伪源域，并在预形状空间中将其特征结构与目标域对齐，从而训练一个能够生成高质量特征的生成器，用于后续的小样本分类等任务。

本章提出的基于预形状空间中测地曲面信息迁移（Information Transfer from the Built Geodesic Surface，ITBGS）的小样本图像生成方法，其整体流程如图 3.1 所示，旨在通过构建一个信息更丰富的伪源域并将其知识迁移至生成器，以克服小样本限制。首先，在伪源域构建阶段，利用少量训练样本，通过 GAN 的判别器提取其深度特征，并将其投影至预形状空间 (Pre-Shape Space)。随后，在预形状空间中根据投

影后的特征点构建测地曲面 (Geodesic Surface)，通过在测地曲面上进行采样或插值，生成大量增强后的特征点，这些增强特征共同构成了伪源域 (Pseudo-source Domain)。与此同时，生成器则以随机隐向量（Latent）为起点，通过在隐空间 (Latent Space) 对这些向量进行插值，生成多样化的输入编码，驱动生成器输出相应的生成图像。为了将伪源域包含的丰富信息迁移给生成器，生成的图像同样通过判别器提取生成特征并投影至预形状空间。核心的信息迁移机制在于通过一个适配损失来对齐伪源域中的增强特征与投影后的生成特征，迫使生成特征去拟合伪源域的分布，从而间接指导生成器学习伪源域所蕴含的知识。此外，为进一步稳定训练并提升生成质量，本章方法还整合了其他辅助约束与监督模块，如图中所示的插值监督和距离正则化约束，它们与适配损失及 GAN 对抗损失共同构成了最终的整体损失函数。本章后续将对提出的伪源域构建方法以及伪源域到目标域的信息迁移方法、插值监督与正则化模块以及最终使用的整体损失函数进行介绍。

3.1.1 基于测地曲面的伪源域构建

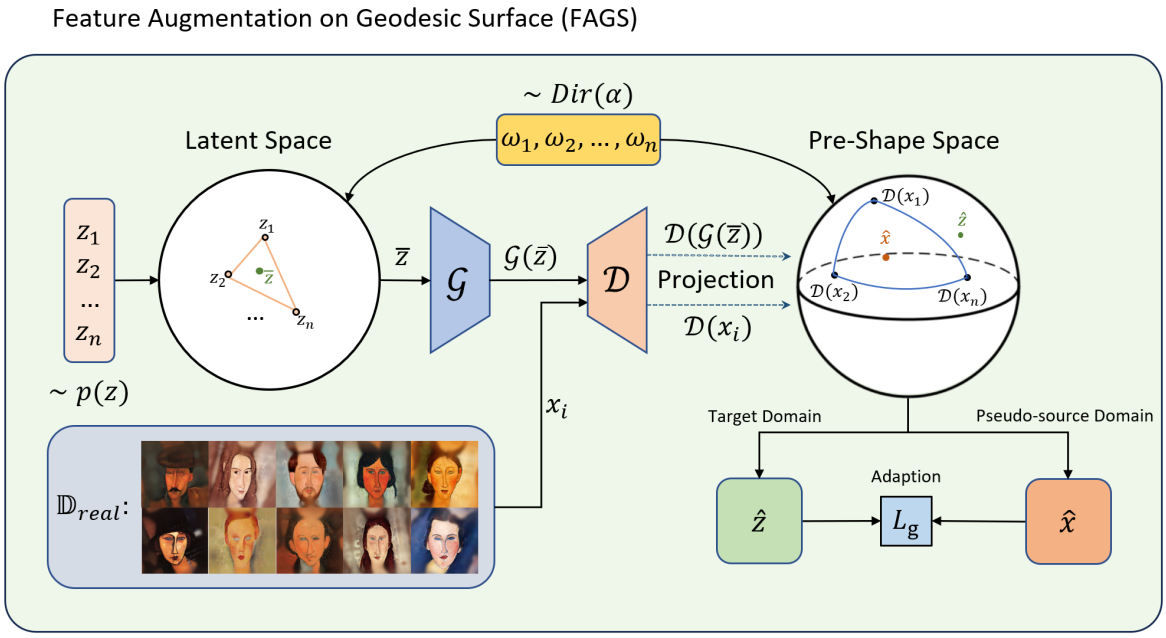


图 3.2 测地曲面增强模块框架图。该模块的核心思想在于，通过约束两种特征，即由锚点隐变量 $\bar{z}$ 生成的特征与在测地曲面上组合真实特征得到的新特征 $\hat{x}$，使得特征之间各自内部的自相关性保持一致，从而提升生成特征的真实性和多样性。

本章通过特征增强构建伪源域，并将伪源域的固有结构信息迁移至目标域，由

此提出测地曲面特征增强模块（Feature Augmentation on Geodesic Surface，FAGS），其算法框架如图 3.2 所示。

特征增强首先需要实现特征提取，鉴于本章聚焦小样本图像生成任务，直接选用图像生成模型进行特征的提取。具体来说，在训练阶段的每一轮中，将训练集 $\mathbb{D}_{real}$ 中的真实图像 $x \sim \mathbb{D}_{real}$ 输入判别器 $\mathcal{D}$ 的第 l 层进行特征提取，得到 $\mathcal{D}^l(x)$，随后将特征投影至预形状空间，用于构建测地曲面。

本论文提出在预形状空间中进行特征增强的核心动机，是将深度神经网络提取的图像特征抽象化为一种形状，并利用形状空间理论挖掘其深层结构信息。具体来说，通过将图像特征投影到预形状空间，实质上引入了一定的先验知识，即特征元素间的相对构型，独立于整体平移和缩放，蕴含了关键的判别性信息[19]。预形状空间分析投影操作移除了这些全局变换的影响，从而更聚焦于特征中更高层次内在结构的表示。与直接在原始高维欧氏特征空间中进行例如线性平均等插值操作相比，基于预形状空间的方法具有显著优势。欧氏插值可能会混合无关的变换信息，导致生成的特征模糊或偏离数据真实分布。相反，在预形状空间中沿测地曲线或在测地曲面上进行插值，是沿着流形上的最短路径或自然路径进行变换[82]。这确保了生成的增强特征是连贯和有效的。它们位于数据的固有流形上，平滑地表示了从一种特征结构到另一种的过渡，并保持了有意义的结构变化[26]。这种基于流形的增强方式适用于小样本学习，因为它能从有限的样本中生成更多样化且结构合理的伪特征，有效扩充数据集，提升模型对数据内在变异模式的学习能力和泛化性能。

传统的测地曲面构建依赖切空间投影，但该过程容易引入重构误差[24]，因此本方法参照了另一种定义方式，将预形状空间中的测地曲面 $\mathbb{G}_{FBS}(\tau, \omega)$ 定义为 Fréchet 重心子空间（Fréchet Barycentric Subspaces，FBS）[83]。由于使用该定义构建测地曲面生成新数据点的过程较为复杂，本论文采用一种等效的方法，通过依次计算测地曲线[22] 上的点来逼近最终的重心，从而生成代表测地曲面上的点。

具体来说，为了生成一个增强特征，我们首先从训练集中选取 n 个真实样本 $\{x_1, \ldots, x_n\}$。提取它们在判别器第 l 层的特征 $\{\mathcal{D}^l(x_1), \ldots, \mathcal{D}^l(x_n)\}$，并通过投影函数 f_p 得到对应的预形状向量集合 $\tau = \{\tau_1, \ldots, \tau_n\}$，其中 $\tau_i = f_p(\mathcal{D}^l(x_i))$。

此迭代计算需要定义权重 ω。在本章方法中，权重向量 $\omega \triangleq \{\omega_1, \ldots, \omega_n\}$ 通过从对称狄利克雷分布 $\mathrm{Dir}(\alpha)$ 中采样获得。这里，n 即为用于生成单个增强特征所选取的输入预形状向量的数量，也对应于狄利克雷分布的维度。我们将所有浓度参数均

设为 $\alpha_i = 1\ (i = 1, \ldots, n)$[49]，这意味着平均而言每个输入预形状 τ_i 对最终生成的点贡献是均等的。

定义好权重后，迭代过程使用如公式 2.5 所示的测地曲线 $\mathbb{G}_{\text{cur}}(\cdot)$，令 μ_j 代表第 j 步迭代得到的中间预形状向量，其中 $j = 1, \ldots, n$，且初始条件设为 $\mu_1 = \tau_1$。后续迭代按下式计算：

$$\mu_j = \mathbb{G}_{cur}(\mu_{j-1}, \tau_j)\left(\frac{\omega_j}{\sum_{i=1}^{j}\omega_i}\right), \quad j = 2, ..., n \tag{3.1}$$

当迭代至 $j = n$ 时，测地曲面上的一个点 $\mathbb{G}_{surf}(\cdot)$ 可以通过一组向量 $\tau = \{\tau_1, ..., \tau_n\}$ 和一组权重 $\omega = \{\omega_1, ..., \omega_n\}$ 得到，表示为：

$$\mathbb{G}_{surf}(\tau, \omega) = \mu_n. \tag{3.2}$$

对于判别器 $\mathcal{D}$ 第 l 层提取的特征张量 $\mathcal{D}^l(x) \in \mathbb{R}^{c\times h\times w}$，为了将其适配于处理二维点集的预形状空间分析框架，使用标准的 PyTorch reshape 操作[109] 将其转换为一个 $2 \times (chw/2)$ 的矩阵，记为 $\mathcal{R}(\cdot)$。具体实现上，这等同于首先将 $c \times h \times w$ 的张量按特定顺序展平成一个长度为 chw 的一维向量，然后将其重新塑形为两行，构成 $m = chw/2$ 个二维点。此操作要求 $c \times h \times w$ 的总元素数为偶数。引入 $\mathcal{R}(\cdot)$ 的目的是将特征图的数值信息强制解释为一种二维空间点集构型，以便应用形状分析理论。

经过维度重塑得到 m 个二维点组成的矩阵后，应用预形状空间投影的标准步骤。结合公式 (2.1) 和 (2.2) 定义的均值消减 $\mathcal{Q}(\cdot)$ 与归一化 $\mathcal{V}(\cdot)$ 操作，最终定义了从原始特征图到其预形状空间表示的完整投影函数 $f_p(\cdot) = \mathcal{V}(\mathcal{Q}(\mathcal{R}(\cdot)))$。通过 $f_p(\mathcal{D}^l(x_i))$ 得到 n 个输入特征的预形状表示 $\{\tau_i\}_{i=1}^n$ 后，便可通过 $\mathbb{G}_{surf}(\tau, \omega)$ 构建测地曲面并生成增强特征的预形状表示 $\hat{x}^l$。将这些生成的增强特征向量 $\hat{x}^l$ 用于构建伪源域 $\mathbb{D}_{ps}$，后续将其中的信息迁移至目标域。

3.1.2 伪源域到目标域的信息迁移

为了将信息从伪源域 D_{ps} 有效迁移至生成器的输出目标域 D_t，需要在两者之间建立一个基于相同生成逻辑的对应关系，并构建一个损失函数进行约束，如图 3.2 所示。上一小节中，伪源域中的每一个增强特征 $\hat{x}^l$ 都是由 n 个真实样本的预形状特征 $\tau = \{\tau_1, \ldots, \tau_n\}$ 和一组权重 $\omega = \{\omega_1, \ldots, \omega_n\}$ 通过构建测地曲面 $\mathbb{G}_{\text{surf}}(\tau, \omega)$ 生成的。为了在目标域中创建一个与 $\hat{x}^l$ 相对应的特征表示 $\hat{z}^l$，采用了在生成器的输入隐空间

中模拟类似组合过程的策略。具体而言，首先随机采样与真实特征数量相同的 n 个隐向量，构成集合 $\{z_i\}_{i=1}^{n}$。然后，使用与生成 $\hat{x}^l$ 时完全相同的权重 ω，通过线性加权平均来计算一个锚定隐向量（Anchor Latent）$\bar{z}$[49]。这个 $\bar{z}$ 旨在作为隐空间中对应于伪源域点 $\hat{x}^l$ 的锚点，其计算公式如下：

$$\bar{z} = \sum_{i=1}^{n} \omega_i z_i, \tag{3.3}$$

将 $\bar{z}$ 输入生成器 $\mathcal{G}$ 得到锚定图像 $\mathcal{G}(\bar{z})$。随后，将 $\mathcal{G}(\bar{z})$ 输入判别器 $\mathcal{D}$ 并通过投影函数 f_p 提取其第 l 层特征的预形状表示 $\hat{z}^l = f_p(\mathcal{D}^l(\mathcal{G}(\bar{z})))$。这些特征向量 $\hat{z}^l$ 构成了目标域 D_t 的表示。

为了将信息从伪源域 D_{ps} 有效迁移至目标域 D_t，期望两域特征在预形状空间中展现出一致的内在结构关系。因此，本论文提出了测地自相关一致性损失（Geodesic Self-correlation Consistency Loss, L_g）。直接计算整个特征图上所有空间位置对之间的自相关性虽然能提供全局结构信息，但对于具有较高空间分辨率 $h \times w$ 的特征图而言，其计算开销和显存占用巨大。

为确保计算可行性，我们采用了一种基于局部块的策略来计算和约束自相关一致性。首先，将在预形状空间中得到的增强特征 $\hat{x}^l$ 和目标特征 $\hat{z}^l$ 重塑回它们各自对应的 $c \times h \times w$ 空间特征图结构，记为 $\hat{x}^l_{\text{map}}$ 和 $\hat{z}^l_{\text{map}}$。为了聚焦于更鲁棒、分辨率更低的特征并进一步提升计算效率，先对这些特征图应用一个自适应平均池化层（AdaptiveAvgPool2d）[109] 来降低空间维度。随后，从每个特征图的不同区域提取并处理 P 个局部块，以捕捉不同位置的结构信息。记这些处理后的块集合为 $\{\hat{x}^l_{\text{map},p}\}_{p=1}^{P}$ 和 $\{\hat{z}^l_{\text{map},p}\}_{p=1}^{P}$，其中每个块 p 的维度为 $c \times K \times K$。

接着，计算每个处理后的局部块 p 内部的块内自相关矩阵（Intra-patch Self-correlation Matrix）。令 $\hat{x}^l_{\text{map},p}(u)$ 表示块 p 中位置 u 处的 c 维特征向量，其中 u 是块内 K^2 个空间位置之一。块内任意两个位置 u, v 之间的自相关 $S^{\hat{x}^l,p}_{u,v}$ 通过它们的余弦相似度进行计算：

$$S^{\hat{x}^l,p}_{u,v} = \langle \hat{x}^l_{\text{map},p}(u), \hat{x}^l_{\text{map},p}(v) \rangle, \tag{3.4}$$

对每个块 p 计算所有 u, v 对即可得到一个 $K^2 \times K^2$ 的相似度矩阵 $S^{\hat{x}^l,p}$。该计算可通过批量矩阵乘法高效完成。

同理，计算目标域特征图 $\hat{z}^l_{\text{map}}$ 中对应每个块 p 的块内自相关矩阵 $S^{\hat{z}^l,p}$：

$$S^{\hat{z}^l,p}_{u,v} = \langle \hat{z}^l_{\text{map},p}(u), \hat{z}^l_{\text{map},p}(v) \rangle. \tag{3.5}$$

最终，测地自相关一致性损失 L_g 通过最小化伪源域和目标域对应块的自相关矩阵之间的差异来实现，从而在局部层面强制结构一致性：

$$L_g = \mathbb{E}_{\substack{z\sim p(z), x\sim \mathbb{D}_{\text{real}}, \\ \omega\sim \text{Dir}}} \sum_l \sum_{p=1}^{P} L_{s\ell_1}(S^{\hat{x}^l,p}, S^{\hat{z}^l,p}), \tag{3.6}$$

其中 l 遍历选定的特征提取器卷积层，p 遍历 P 个处理的局部块，$L_{s\ell_1}(\cdot)$ 代表逐元素应用的 smooth-ℓ_1 损失函数[110]。这种基于局部块的计算方式相比全局计算，显著降低了计算量和显存需求，同时仍能有效地捕捉并迁移局部结构信息。

3.1.3 插值监督与正则化模块

对于生成模型 $\mathcal{G}$ 性能的评估，一个常用的方法是考察其在隐向量空间进行插值时生成图像序列的质量[49]。该过程首先随机采样两个隐向量 $z'_1, z'_k \sim p(z)$，然后在它们之间定义一个线性插值路径，生成一系列中间隐向量 $Z_{inp}(z'_1, z'_k) = \{z'_1 + \frac{i}{k}(z'_k - z'_1)\}^k_{i=0}$。将这个序列 Z_{inp} 输入生成器 $\mathcal{G}$，即可获得相应的插值图像序列 $\mathcal{G}(Z_{inp})$。理想情况下，这个图像序列应当展现出平滑、自然的视觉过渡。

图 3.3 结合 FAGS 模块的 StyleGAN2 插值生成图像

然而，在小样本场景下训练的模型，观察到生成的中间插值图像可能存在模糊或不连续的阶梯状变化（Stair-like Phenomenon），如图 3.3 所示。这种情况往往表明生成器可能对训练数据产生了过拟合或记忆，未能学习到真正平滑且泛化的数据流形[49]。为了解决并改善这种在隐空间插值过程中生成图像的平滑性和连续性问题，本小节提出图 3.4 所示的插值监督与正则化（Interpolation and Regularization，I&R）模块。该模块旨在通过一种双阶段优化策略，结合算法 3.1 的流程，提升模型生成平滑且多样化插值序列的能力。

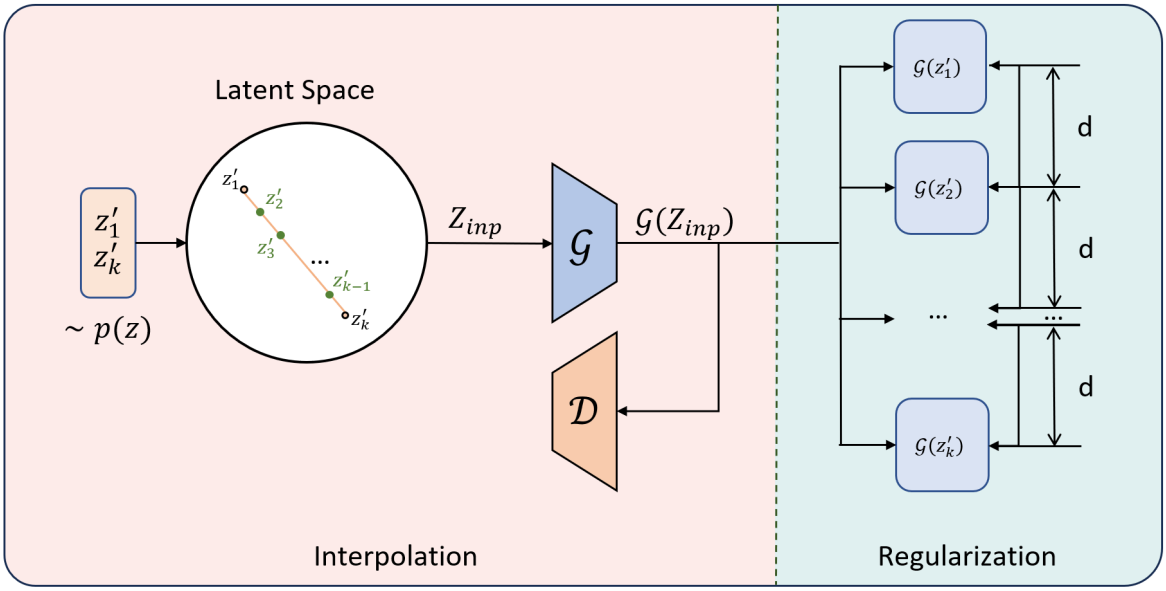


图 3.4 插值监督与正则化（Interpolation and Regularization，I&R）模块架构示意图

算法 3.1: 插值监督与正则化模块 (I&R) 的算法流程

Require: z1, zk: 随机采样的起始与结束隐向量
Require: k: 插值样本数

```
# Interpolation (插值监督阶段)
Z_inp = cat([lerp(z1, zk, v) for v in linspace(0, 1, k)])   ▷ 生成插值隐向量序列 Z_inp
inp_imgs, inp_feats = Generator(Z_inp)                      ▷ 生成器 G 生成插值图像及其中间特征
# inp_feats : k × c × h × w                                 ▷ 中间特征图的维度示例
pred = Discriminator(inp_imgs)                              ▷ 判别器 D 对插值图像序列进行评估
L_inp = log(pred).mean()                                    ▷ 对应公式 3.7

# Regularization (特征距离正则化阶段)
dist() = L2_distance()                                      ▷ 定义特征间距离的度量方式
inp_feats = AdaptiveAvgPool2d(inp_feats)                    ▷ 应用自适应平均池化
# k × c × h × w → k × c × (h/4) × (w/4)                     ▷ 池化后特征图的维度示例
inp_feats_temp = cat([inp_feats[1 :], inp_feats[0]])        ▷ 创建特征序列的循环移位版本
feats_dist = dist(inp_feats, inp_feats_temp)                ▷ 计算插值序列中相邻特征之间的距离
q_dist = norm(cat([ones(k − 1), Tensor([k − 1])]))          ▷ 构建目标分布 Q
L_dr = KLDivLoss(log_softmax(feats_dist), q_dist)           ▷ 计算 KL 散度损失 L_dr
```

对于插值监督模块，将生成器产生的完整插值图像序列 $\mathcal{G}(Z_{inp})$ 输入判别器 $\mathcal{D}$，并采用对插值图像的判别损失 L_{inp} 进行优化，如下所示：

$$L_{inp} = \mathbb{E}_{z'_1, z'_k \sim p(z)} \left[\log \mathcal{D}(\mathcal{G}(Z_{inp}(z'_1, z'_k))) \right], \tag{3.7}$$

该损失鼓励生成器 $\mathcal{G}$ 学习一个更平滑的隐空间流形，使得路径上的所有插值点都能生成逼真的图像。

为了进一步消除阶梯状伪影并提升插值平滑性，引入了特征距离正则化项 L_{dr}，

以引导生成器 $\mathcal{G}$ 学习更平滑的特征流形。为了计算 L_{dr}，提取中间隐向量序列 Z_{inp} 通过生成器 $\mathcal{G}$ 后的对应中间特征，并进行池化处理。接着，计算这些处理后特征序列中循环相邻特征之间的 ℓ_2 距离 d。这些距离随后通过 Softmax 函数转换为实际的概率分布 $P(d)$。此分布将与一个特殊设计的目标概率分布 Q 进行比较。目标分布 Q 的具体形式和构建方式遵循算法 3.1 中的定义，其设计旨在鼓励插值路径中间步骤的特征距离保持一致，同时确保路径首尾特征之间存在显著差异。特征距离正则化损失 L_{dr} 即为这两个分布之间的 KL 散度：

$$L_{dr} = \mathbb{E}_{z'_1, z'_k \sim p(z)} \left[D_{KL}(Q || P(d)) \right], \tag{3.8}$$

最小化此损失会促使生成器产生的特征距离分布 $P(d)$ 趋向于目标分布 Q，从而改善插值的平滑性和连续性。

3.1.4 最终优化目标函数

本章提出了基于预形状空间中测地曲面信息迁移的小样本图像生成方法，包含 FAGS 模块和 I&R 模块，在本节定义生成器 $\mathcal{G}$ 与判别器 $\mathcal{D}$ 的联合优化目标。生成器的复合损失函数 $L^{\mathcal{G}}$ 可表述为：

$$L^{\mathcal{G}} = L^{\mathcal{G}}_{adv} - \lambda_1 L_{inp} + \lambda_2 L_{dr}, \tag{3.9}$$

而判别器的优化目标则定义为：

$$L^{\mathcal{D}} = L^{\mathcal{D}}_{adv} + \lambda_1 L_{inp} + \lambda_3 L_g, \tag{3.10}$$

其中，$L^{\mathcal{G}}_{adv}$ 与 $L^{\mathcal{D}}_{adv}$ 为经典对抗训练损失项，其数学形式如式 (2.7) 与 (2.8) 所示，L_{inp} 为插值监督损失项，同时在生成器与判别器更新过程中起约束作用，L_{dr} 为特征距离正则化项，专用于生成器参数更新。L_g 为测地自相关一致性损失，专用于判别器更新，这种做法增加了判别器任务的复杂度，起到正则化判别器的作用，从而为生成器提供更稳定和持续的学习信号[4]。λ_1, λ_2 和 λ_3 为预设平衡系数，用于调节各损失项的贡献权重。

3.2 实验分析

本节首先介绍实验设置和数据集，接着，展示在多种数据集上的实验结果，包括定性对比实验、定量对比实验和消融实验，最后在特定领域的小样本数据集上验证使用本章方法提出的生成模型所生成图像的有效性。

3.2.1 实验设置

本节实验的框架基于 StyleGAN2 架构[90] 构建，并融合 MixDL[49] 模块。为应对极端小样本场景，实验中全程禁用了自适应数据增强 (Adaptive Data Augmentation, ADA)[4] 策略。损失函数 (3.9) 与 (3.10) 中的平衡系数为 $\lambda_1 = 0.8$、$\lambda_2 = 1.25$ 和 $\lambda_3 = 0.8$。训练的批量规模与插值样本数均设为 4。这些基础设置，结合为测地自相关一致性损失 L_g 所采用的基于局部块的高效计算方法，共同确保了整个训练过程能在单张 NVIDIA GeForce RTX 3090 GPU（24GB 显存）平台上稳定运行。

具体在计算 L_g 时，采用了如下策略以降低计算和显存负载：首先对相关的特征图 $\hat{x}^l_{\text{map}}$ 和 $\hat{z}^l_{\text{map}}$ 应用一个自适应平均池化层，将它们的空间维度从 $h \times w$ 降采样至 $(h/2) \times (w/2)$。随后，在这些经过池化的特征图上，选取四个象限和中心区域共 $P = 5$ 个代表性位置，分别提取并处理 P 个 $K \times K$ 大小的局部块以计算块内自相关矩阵。其中，块的边长 K 根据池化后的特征图高度 $h_{\text{pooled}} = h/2$ 设定，具体取 $K = h_{\text{pooled}}/2 = h/4$。

实验选取 N-div[111]、MSGAN[112]、DistanceGAN[113]，以及 StyleGAN2（SG2）[90]、StyleGAN2+ADA（SG2A）[4]、FastGAN（FG）[43]、PatchDiffusion（PD）[44] 和 MixDL（MDL）[49] 进行定性与定量对比分析，其中 MixDL 作为当前无需源域的小样本生成最优方法，是本章主要的对比基准方法。

数据集涵盖艺术创作（Amedeo Modigliani 绘画[114]、风景素描[7]）、小狗脸部（Animal-Face Dog[115]）、绘制图像（动漫人脸[43]、Pokemon[43]、人脸素描[116]）、真实人像（FFHQ[89]、CelebA[117]）及材料科学图像（UHCSDB 钢铁材料显微图像[118] 和铜合金微观图像（CuCrZr））等多种类图像数据。对于 UHCSDB 数据集，挑选了其中碳化物网络（Carbide Network）和球状体（Spheroidite）类别中的图片用于生成图像的定性与定量实验分析。对于 CuCrZr 数据集，主要用于生成图像有效性的分析，将会在小节 3.2.6 中详细介绍。其中 Amedeo Modigliani 绘画和风景素描数据集均只有 10 张图像，对于其余图像数量大于 10 张的数据集，随机选取其中 9 张或 10 张图像，

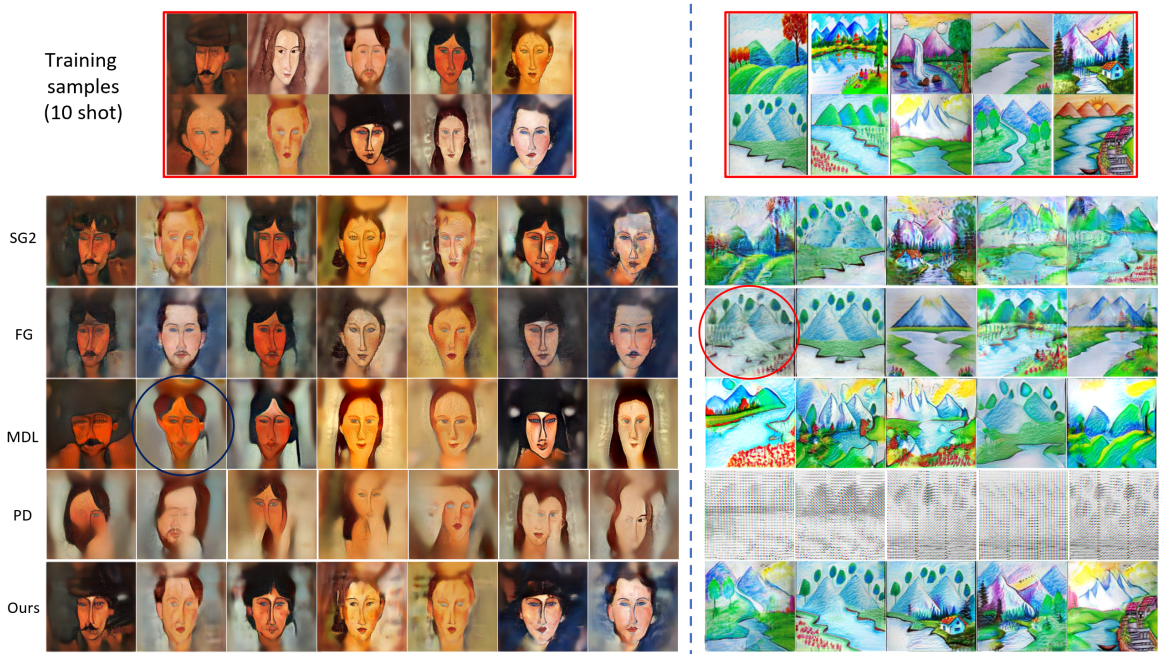


图 3.5　在包含 10 个样本的 Amedeo Modigliani 绘画（左）和风景素描（右）数据集上各种方法的生成图像结果。SG2 表示 StyleGAN2、FG 表示 FastGAN、MDL 表示 MixDL、PD 表示 PatchDiffusion、Ours 表示本章方法。

作为子集用于极小样本图像生成任务的训练集。实验用到的所有图片的分辨率均为 256×256。

3.2.2　定性对比实验

图 3.5 展示了在 Amedeo Modigliani 绘画与风景素描这两个数据集上，不同方法的生成结果。值得注意的是，除了 FastGAN[43] 在计算感知损失[106] 时使用了一个预训练的 VGG 之外，其余方法皆从头进行训练，且没有使用任何额外的辅助信息。从图 3.5 可以看出，StyleGAN2[90] 在两个小样本数据集上生成的图像均出现了模糊或过拟合的情况。FastGAN[43] 在 Amedeo Modigliani 绘画上取得了与本章方法相当的效果，但在风景素描数据集上的效果不佳，甚至只是把训练集中第二行的前两张图像做了简单的加权叠加，见图 3.5 中红色圈出的生成结果。PatchDiffusion[44] 生成的图像则像是训练集中图像元素的生硬拼接，视觉质量较低。MixDL[49] 在多样性方面表现突出，但在保真度上不及其他方法，如图 3.5 中蓝色圈出的示例，其人脸形状存在明显变形。值得关注的是，本章方法在两个小样本数据集上的生成图像都展现出了良好的保真度和多样性。这主要得益于对真实图像中形状、颜色、纹理等视觉元素的自然融合。基于数据集中提取的图像特征投影至预形状空间中构建测地曲面，所

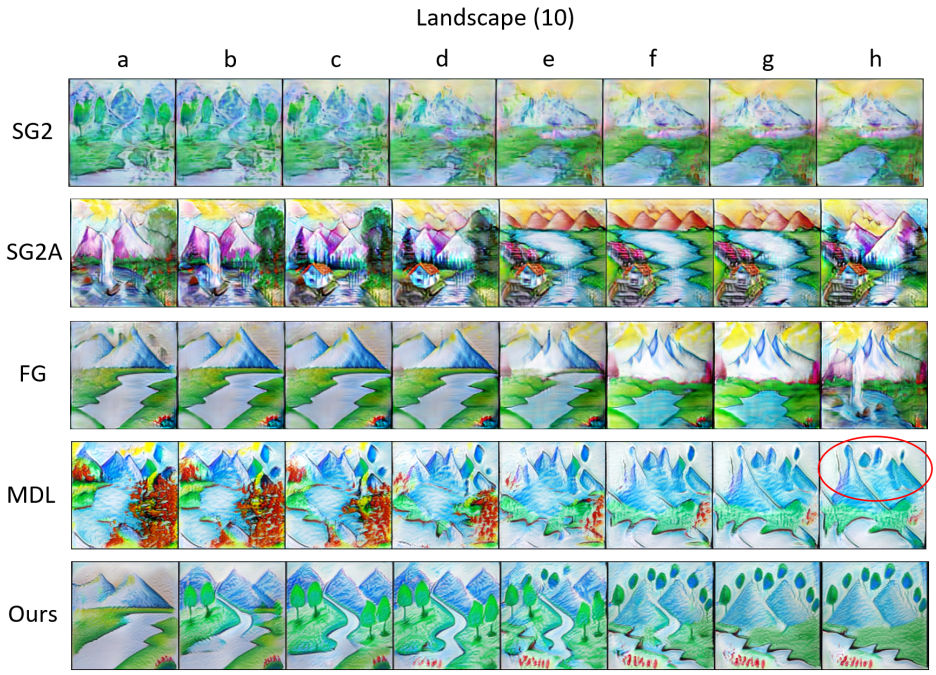


图 3.6　在包含 10 个样本的风景素描数据集上的插值图像生成结果。SG2 表示 StyleGAN2、SG2A 表示 StyleGAN2+ADA、FG 表示 FastGAN、MDL 表示 MixDL、Ours 表示本章方法。

生成的特征能够自然地整合多张图像中的视觉元素，从而帮助模型生成更自然融合的结果。

图 3.6 、图 3.8和图 3.7 展示了部分方法在多组小样本数据集上生成的插值图像。如图 3.7 中列 b 到列 c 的所示，FastGAN 在动漫人脸数据集中出现了明显的阶梯状变化现象，同时列 c 和列 d 也能观察到 FastGAN 生成图像的缺陷，但其在人脸素描数据集上有不错的插值结果。MixDL[49] 虽然能生成平滑的语义插值，但保真度有所下降，表现为在图 3.6 中红色圈出的山峰部分看起来较为怪异。StyleGAN2+ADA[4] 在动漫人脸数据集上也存在类似的保真度与多样性平衡问题，而原始的 StyleGAN2[90] 虽能生成具有较好保真度的插值图像，但从列 e 到列 f 依然会出现阶梯状变化现象，这一点在图 3.7 和图 3.8 都有所体现。本章方法则可以在各类仅包含 10 个样本的数据集中进行平滑的潜变量插值，同时保持足够的保真度，这也再次说明了本章方法在平衡保真度和多样性方面的有效性。

在真实人脸数据集上的生成结果可作为衡量生成模型质量的关键标准。图 3.9 展示了在仅有 9 个样本的 FFHQ[89] 子集上的实验结果，可以看出本章方法能够将两张或多张人脸的特征自然结合。尤其是在发型、胡子等面部细节的平滑融合上表现突出。值得注意的是，FastGAN[43] 在 FFHQ 数据集上的结果同样比较出色，而其他对

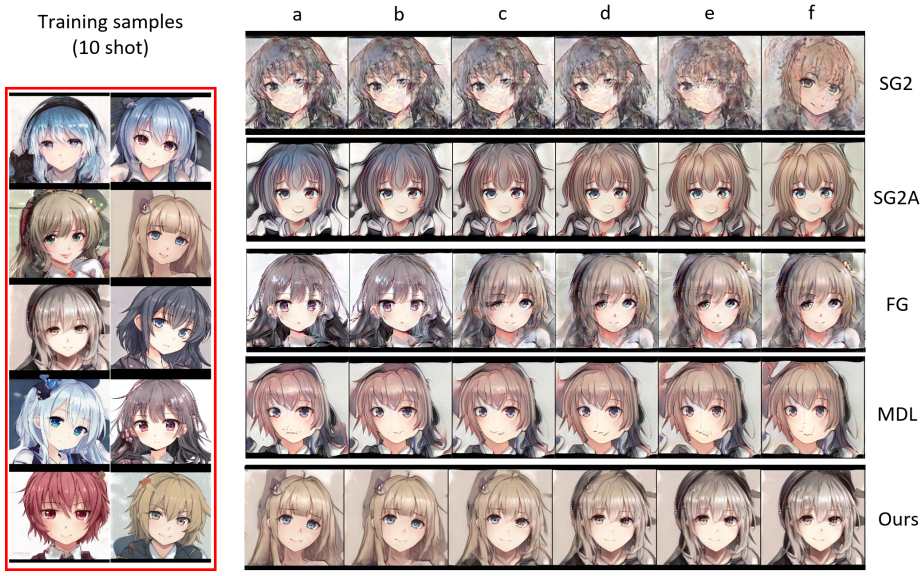


图 3.7 在包含 10 个样本的动漫人脸数据集上的插值图像生成结果。SG2 表示 StyleGAN2、SG2A 表示 StyleGAN2+ADA、FG 表示 FastGAN、MDL 表示 MixDL、Ours 表示本章方法。

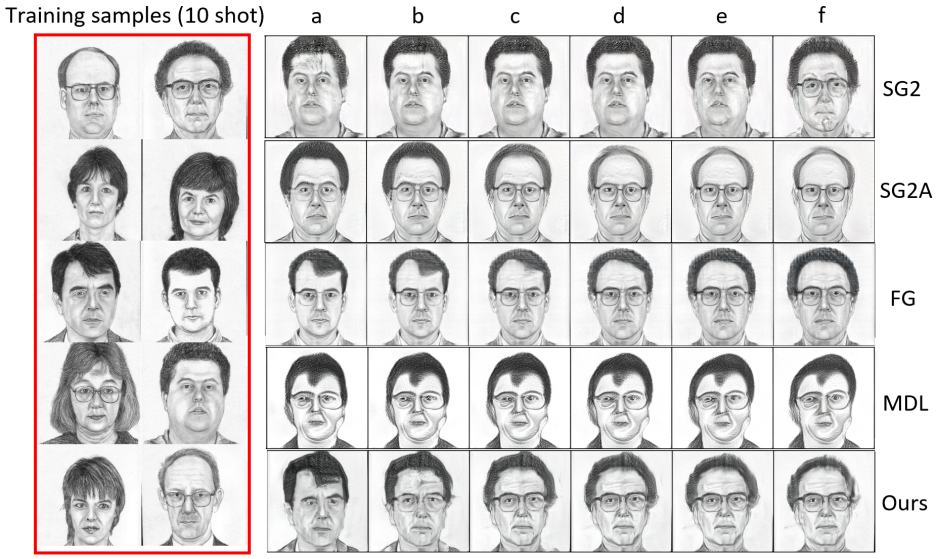


图 3.8 在包含 10 个样本的人脸素描数据集上的插值图像生成结果。SG2 表示 StyleGAN2、SG2A 表示 StyleGAN2+ADA、FG 表示 FastGAN、MDL 表示 MixDL、Ours 表示本章方法。

比方法则难以生成逼真的人脸图像。

在材料领域，由于样本采集难度较大，小样本问题为后续下游任务带来了巨大挑战。通过训练小样本图像生成模型进行数据增强，已成为一种潜在的可行方案。为此，从 UHCSDB[118] 数据集中选取了碳化物网络和球状体两个类别，并随机挑选 10 张钢显微结构图像作为训练集。图 3.10 和图 3.11 展示了相应的实验结果。FastGAN、MixDL 和 PatchDiffusion 在生成纹理图像时效果不佳，未能准确反映钢的微观结构。

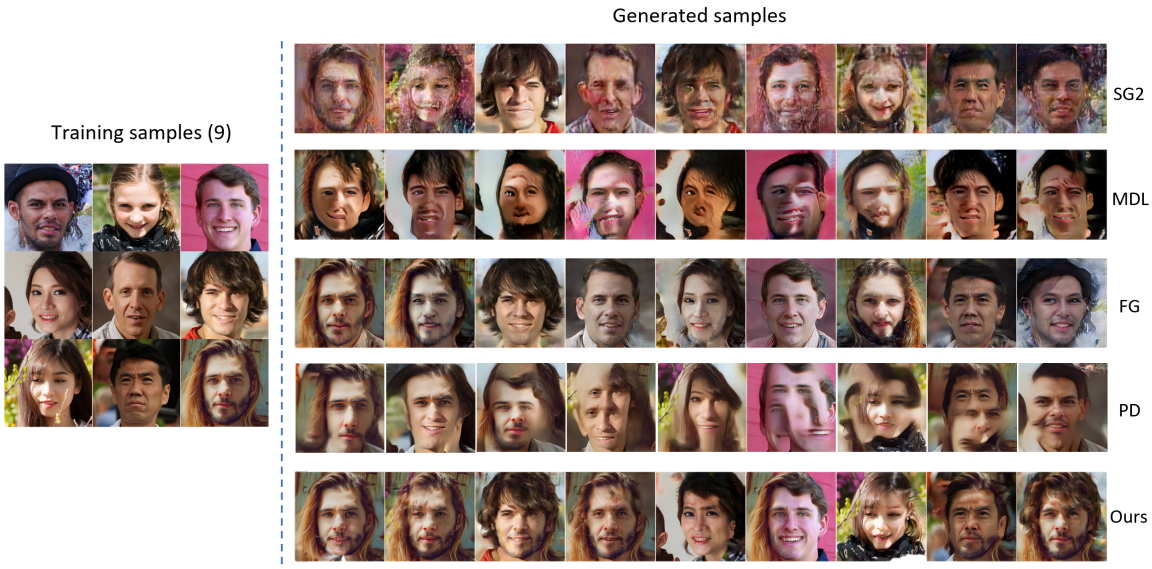


图 3.9 在仅有 9 个样本的 FFHQ 子集上的各个方法的生成图像结果。SG2 表示 StyleGAN2、MDL 表示 MixDL、FG 表示 FastGAN、PD 表示 PatchDiffusion、Ours 表示本章方法。

StyleGAN2 在生成球状体显微图像时出现了过拟合，而 StyleGAN2+ADA 虽能在碳化物网络上取得良好效果，但在球状体数据集上生成的晶粒存在模糊情况。相较之下，本章方法在两种不同的材料显微结构上都表现稳定，具有更高的鲁棒性和可靠性。在后续小节 3.2.6 中，将进一步对小样本材料数据集上本章方法生成的图像进行有效性分析。

3.2.3 定量对比实验

在定量分析中，主要使用 Fréchet Inception Distance (FID)[105] 和 pairwise Learned Perceptual Image Patch Similarity (LPIPS)[106] 两种指标。其中 FID 在小样本数据集上计算，而 LPIPS 则在生成样本之间进行计算。FID 越低、LPIPS 越高分别表示更好的图像质量和更高的多样性。

表 3.1、表 3.2 以及表 3.3 展示了在多个小样本数据集上各方法的定量结果。本章方法在不同数据集上的参数设置保持一致，未进行特定于某一领域的微调。可以看到，评估生成模型能力需要同时考虑生成图像的保真度和多样性，这两个方面分别可以由 FID 和 LPIPS 反映。首先观察表 3.1，本章方法上在 FID 和 LPIPS 上大多取得了最优或接近最优的表现。PatchDiffusion 方法在多数数据集上取得了较高的 LPIPS 分数，表明该方法的多样生成能力，但结合定性实验的生成图像展示来看，该方法生成的图像接近与训练集图像的生硬拼接，与真实图像差异较大，视觉质量不

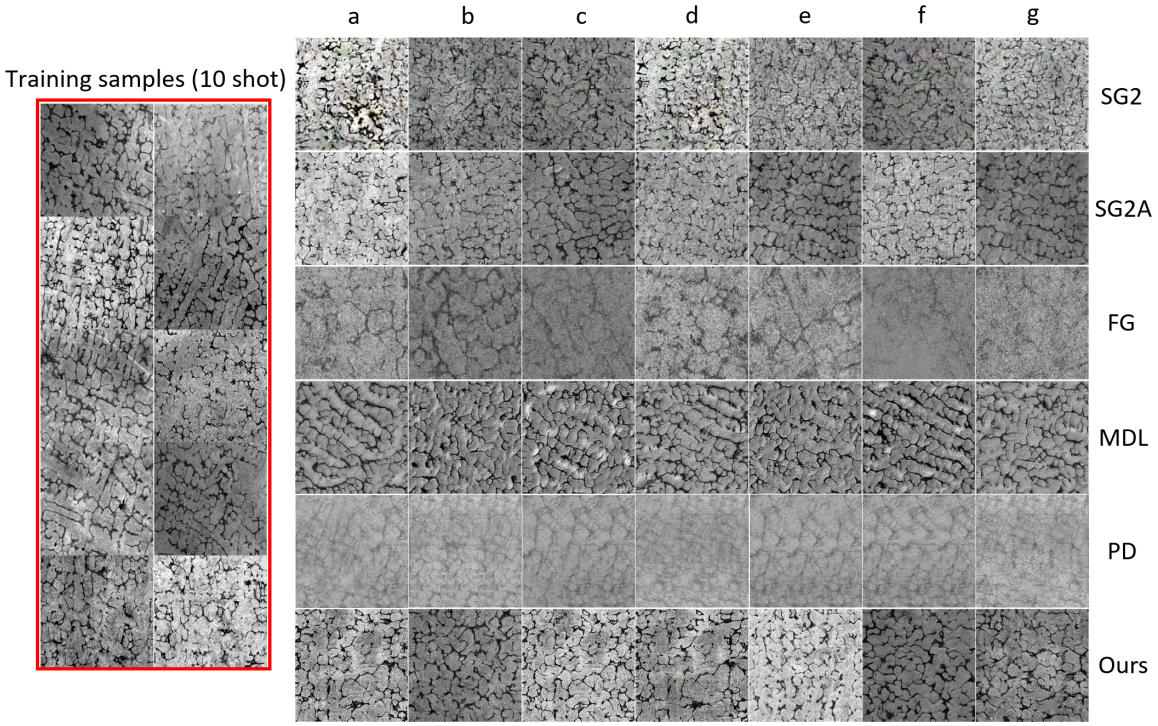


图 3.10　在包含 10 个样本的碳化物网络子集上不同方法的生成图像结果。SG2 表示 StyleGAN2、SG2A 表示 StyleGAN2+ADA、FG 表示 FastGAN、MDL 表示 MixDL、PD 表示 PatchDiffusion、Ours 表示本章方法。

高。尽管本章方法在 Pokemon 和风景素描这两个数据集上的指标略低于 FastGAN[43]，但依然能保持第三好的成绩，最佳的结果由 FastGAN 结合本章提出的 FAGS 模块（FastGAN+FAGS）得到。而在真实人脸数据集上，本章方法和 FastGAN 生成的样本有着相近的视觉质量，如图 3.9 所示，从表 3.2 的量化指标来看，本章方法在多样性方面更具优势。PatchDiffusion 虽然在多样性指标上达到了最佳，但保真度指标较低，生成图像的视觉质量明显不如本章提出的方法与 FastGAN。

在材料显微结构数据集上，图 3.10 和图 3.11 中的结论也得到了表 3.3 的支持。StyleGAN2 和 StyleGAN2+ADA 分别在球状体和碳化物网络数据集上有较低的 FID，说明在保真度方面较为出色。但它们的 LPIPS 得分显著低于本章方法，说明在生成图像的多样性上本章提出的方法更优。MixDL 和 FastGAN 在这两个材料数据集上的 FID 较高。相较之下，本章方法在两项指标上都取得了最好和第二好的成绩。

此外，将 FAGS 模块集成到 FastGAN 中，在 Pokemon、风景素描、CelebA、碳化物网络和球状体等数据集上也能显著提升性能，体现了 FAGS 模块插入到不同生成器架构中的普适性。

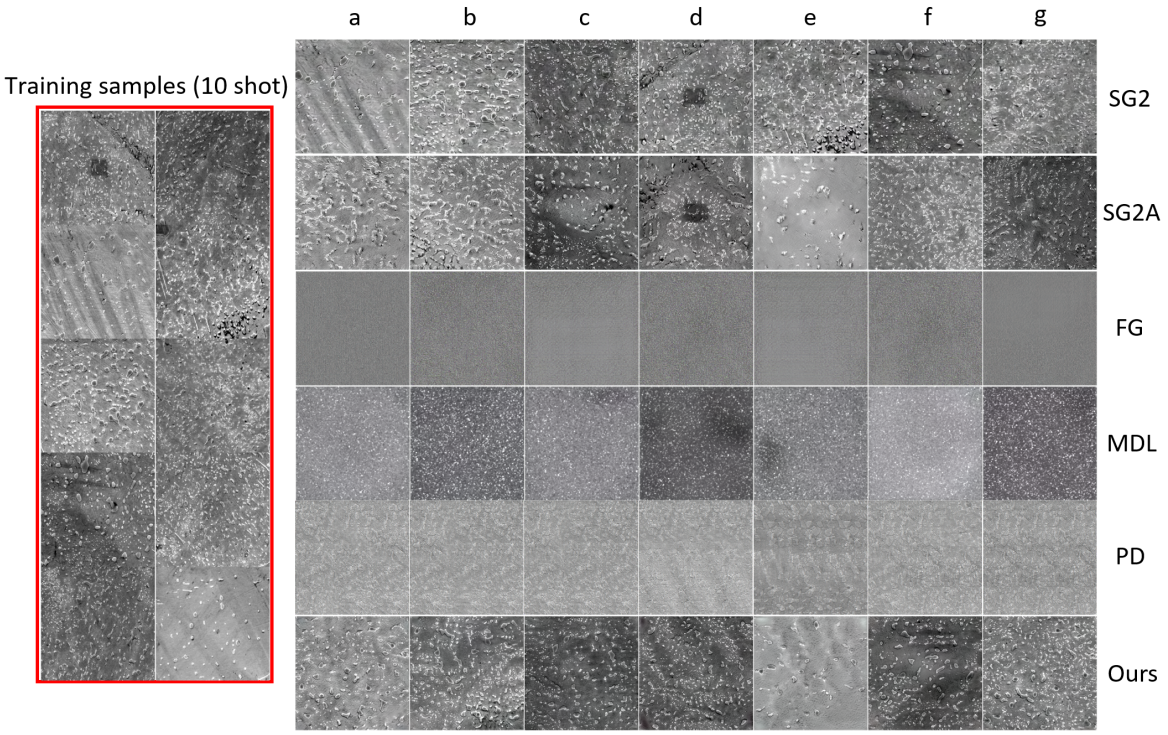


图 3.11 在包含 10 个样本的球状体子集上不同方法的生成图像结果。SG2 表示 StyleGAN2、SG2A 表示 StyleGAN2+ADA、FG 表示 FastGAN、MDL 表示 MixDL、PD 表示 PatchDiffusion、Ours 表示本章方法。

3.2.4 消融实验

为了探究本章提出的各个模块的有效性以及模块内部设计的选择，本章设计了一系列消融实验，具体包括各个模块对于不同生成器架构的影响，特征增强和信息迁移策略的选择，以及插值监督（I）与正则化（R）模块的协同作用。

如表 3.4 所示，本实验使用了 StyleGAN2 与 FastGAN 作为基准架构，评估了各模块的贡献度。对于 StyleGAN2，单独引入 FAGS 模块可使 FID 指标降低 3.4，LPIPS 提升 9.8%。当联合启用插值监督模块后，性能出现显著改善，FID 进一步下降 24.6，LPIPS 提升 4%。最终完整引入正则化模块时取得最优结果，FID 降低 91.6 达 90.7，LPIPS 达到 0.677，验证了本章提出的模块协同作用的有效性。

在 FastGAN 架构中，I&R 模块的引入导致 FID 指标上升 14.9，LPIPS 下降 2%，这表明 I&R 模块的效果可能具有架构依赖性，FastGAN 在仅引入 FAGS 模块的情况下取得最优结果。如图 3.3 的插值图像可视化分析显示，StyleGAN2 在未引入 I&R 模块时出现显著模糊现象，而观察图 3.8，FastGAN 因本身紧凑的架构设计在一些数据集上对中间插值具有更好的鲁棒性。

表 3.1　在小样本图像生成任务上的定量结果。最好和次好的结果分别以粗体和下划线标出，F 和 L 分别表示 FID 和 LPIPS 指标。

方法	动漫人脸		小狗脸部		人脸素描		Amedeo 绘画		风景素描		Pokemon	
	F(↓)	L(↑)	F(↓)	L(↑)	F(↓)	L(↑)	F(↓)	L(↑)	F(↓)	L(↑)	F(↓)	L(↑)
N-Div[111]	175.4	0.425	150.4	0.632	–	–	–	–	–	–	–	–
MSGAN[112]	138.6	0.536	165.7	0.630	–	–	–	–	–	–	–	–
DistanceGAN[113]	84.1	0.543	102.6	0.678	–	–	–	–	–	–	–	–
StyleGAN2[90]	213.9	0.407	312.9	0.549	188.4	0.476	**68.6**	**0.649**	210.3	0.531	261.9	0.475
StyleGAN2+ADA[4]	282.3	0.473	342.0	0.539	341.3	0.469	216.3	0.538	207.7	0.498	278.5	0.413
MixDL[49]	140.9	0.529	291.1	0.701	137.9	0.396	205.2	0.643	183.3	0.698	231.2	0.499
FastGAN[43]	150.5	0.393	65.1	0.671	112.4	0.437	108.3	0.615	83.8	0.689	203.3	**0.554**
PatchDiffusion[44]	132.6	**0.545**	196.0	0.704	94.7	**0.528**	231.8	0.596	344.2	0.387	215.4	0.545
FastGAN+FAGS	123.2	0.304	**54.5**	0.679	97.8	0.292	98.9	0.588	**82.6**	**0.699**	**200.8**	0.420
本章方法	**72.5**	0.538	95.0	**0.713**	**57.7**	0.485	113.9	0.647	90.7	0.677	208.1	0.552

表 3.2　在 FFHQ 和 CelebA 上的定量结果。最好和次好的结果分别以粗体和下划线标出。

方法	FFHQ		CelebA	
	FID(↓)	LPIPS(↑)	FID(↓)	LPIPS(↑)
StyleGAN2	311.6	0.442	102.3	0.561
MixDL	283.7	0.640	206.8	0.531
FastGAN	**112.0**	0.593	86.6	0.507
PatchDiffusion	221.8	**0.642**	183.7	**0.595**
FastGAN+FAGS	220.9	0.448	**67.3**	0.554
本章方法	130.9	0.617	91.3	0.570

为评估本章提出的 FAGS 模块的整体有效性，并分析其内部特征增强和信息迁移策略起到的作用，本节进行了消融实验。图 3.12 展示了在 Amedeo 绘画数据集上不同方法设置下的定性生成结果。图中每行代表一种不同的方法设置，左侧为随机生成图像，右侧标记为 a 至 f 的图像序列则展示了对应的隐空间插值效果。

首先观察图中的第五行，展示了本章提出的完整 FAGS 方法，该方法结合了测地曲面特征增强和测地自相关一致性损失。可以看到，生成的随机样本视觉质量和结构一致性较好，并且插值序列过渡平滑自然，没有明显的伪影或阶梯状变化，验证了完整方法的有效性。作为对比，前三行展示了不使用测地曲面增强，而是基于类似 RSSA[8] 的框架结合不同信息增强或迁移策略的效果。具体地，标记为 RSSA with L_{scc} 的第一行，移除预训练生成模型后，直接将训练图像作为源域并使用原始的自

表 3.3　在碳化物网络和球状体数据集上的定量结果。最好和次好的结果分别以粗体和下划线标出。

方法	碳化物网络		球状体	
	FID(↓)	LPIPS(↑)	FID(↓)	LPIPS(↑)
StyleGAN2	209.4	0.440	**209.5**	0.378
StyleGAN2+ADA	<u>120.4</u>	0.339	276.6	0.422
MixDL	163.3	**0.541**	529.8	0.397
FastGAN	227.2	0.272	583.2	0.282
PatchDiffusion	372.6	0.373	515.6	0.293
FastGAN+FAGS	124.2	0.508	249.7	<u>0.490</u>
本章方法	**118.4**	<u>0.523</u>	<u>248.9</u>	**0.527**

表 3.4　本章提出的各个模块的定量消融实验

生成器架构	FAGS	I	R	风景素描	
				FID(↓)	LPIPS(↑)
StyleGAN2	✗	✗	✗	210.3	0.531
	✓	✗	✗	206.9(-3.4)	0.629(+0.098)
	✓	✓	✗	182.3(-24.6)	0.669(+0.04)
	✓	✓	✓	**90.7(-91.6)**	**0.677(+0.008)**
FastGAN	✗	✗	✗	83.8	0.689
	✓	✗	✗	**82.6(-1.2)**	**0.699(+0.01)**
	✓	✓	✓	97.5(+14.9)	0.679(-0.02)

相关一致性损失（L_{scc}）将其信息迁移到目标生成器。如图所示，由于小样本源域中信息有限，其生成图像和插值图像均显得较为模糊，细节丢失严重。第二行展示了结合经典 Mixup[28] 线性组合策略的效果，即 RSSA with Mixup，由于使用 Mixup 所构建的伪源域可以视为对多张训练样本的线性组合，因而生成的图像存在可见的叠影现象。第三行为结合 Manifold Mixup[10] 的结果，即 RSSA with M Mixup。该方法效果虽略有改善，但图像清晰度和插值平滑性仍有不足。这表明简单的线性组合策略难以有效处理小样本数据的复杂分布。与线性组合方法相比，FAGS 模块通过构建预形状空间上的测地曲面来丰富信息，这种方式更能平滑地融合训练样本中的元素。

此外，标记为 FAGS with $L_{s\ell_1}$ 的第四行保留了 FAGS 的测地曲面特征增强，但将信息迁移损失从测地自相关一致性损失 L_g 替换为简单的 Smooth-ℓ_1 特征匹配损失。比较第四行和第五行可以看出，相较于使用 L_g 的完整 FAGS 方法，使用其他信息迁

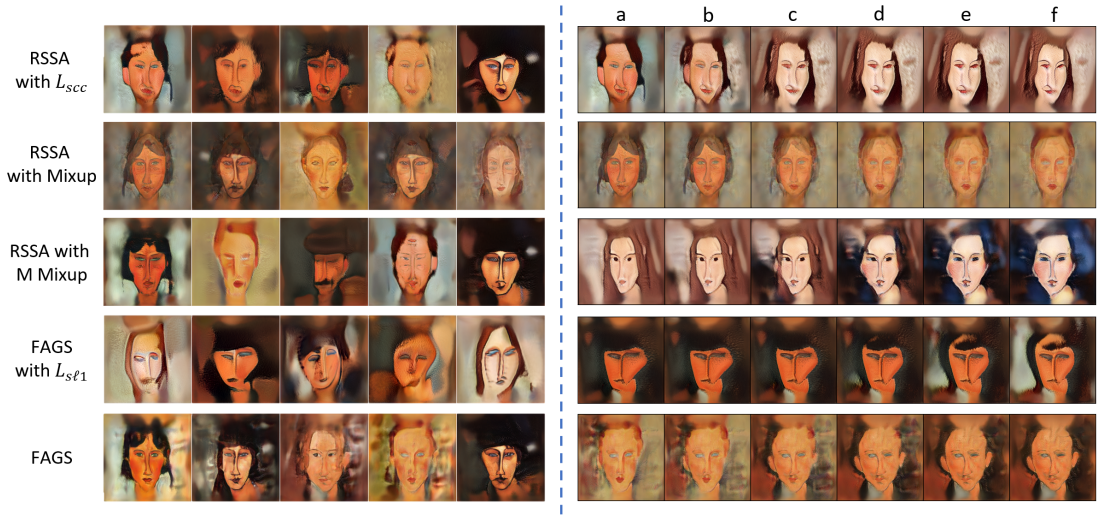


图 3.12　FAGS 模块的定性消融实验对比。左侧为生成图像，右侧为插值图像。

表 3.5　FAGS 模块的定量消融实验对比

方法	FID(↓)	LPIPS(↑)
RSSA+L_{scc}	186.8	0.585
RSSA+Mixup	237.6	0.516
RSSA+Manifold Mixup	199.4	0.630
FAGS+$L_{s\ell 1}$	187.1	0.527
FAGS+L_g	**113.9**	**0.647**

移损失生成图像的结构感和插值序列的连贯性稍逊一筹。这凸显了 L_g 损失在保持和迁移特征内在结构信息方面的重要性。

如表 3.5 所示，对比传统数据增强策略，FAGS 模块展现出显著优势。与 RSSA 基线方法相比，采用测地自相关一致性损失 L_g 的 FAGS 模块将 FID 降低 62.9，LPIPS 提升 10.6%。即便将损失替换为 smooth-ℓ_1 损失，性能仍优于基于 Manifold Mixup 的方法。

如图 3.13 所示，插值监督与正则化（I&R）模块对插值图像的质量具有关键作用。图 3.13 的前两行中，没有使用插值监督与正则化模块的基线模型在 b - e 列的中间插值位置出现显著模糊。引入 L_{inp} 损失的生成结果如第 3 - 4 行所示，生成清晰度提升，模糊问题得到缓解，但观察从列 a 过渡到列 b，以及从列 c 过渡到列 d 时的视觉变化，可以发现插值图像仍然存在阶梯状变化现象。第 5 行与第 6 行展示了完整插值监督与正则化模块的生成结果，视觉质量最佳且插值图像的过渡平滑自然，验证了插值监督与距离正则化策略的有效性。

图 3.13　插值监督与正则化模块的定性消融实验对比。前两行为没有添加插值监督与正则化模块的结果、第 3 - 4 行为仅添加插值监督模块，未添加正则化模块的结果、最后两行为两个模块均添加的结果。

表 3.6　风景素描数据集上的参数敏感性分析。本章方法配置为：$\lambda_1 = 0.8$、$\lambda_2 = 1.25$、$\lambda_3 = 0.8$。最优结果已加粗显示。

配置	λ_1	λ_2	λ_3	FID ↓	LPIPS ↑
降低 λ_3	0.8	1.25	0.4	211.17	0.616
降低 λ_1	0.4	1.25	0.8	202.39	0.604
降低 λ_2	0.8	0.25	0.8	206.32	0.566
本章方法	0.8	1.25	0.8	**90.70**	**0.677**

如公式 (3.9) 和 (3.10) 所示的最终优化目标利用了三个关键超参数 λ_1、λ_2 和 λ_3，分别用以平衡插值损失 L_{inp}、距离正则化 L_{dr} 以及测地线自相关一致性损失 L_g。本章方法的标准配置将这些参数设定为 $\lambda_1 = 0.8$、$\lambda_2 = 1.25$ 和 $\lambda_3 = 0.8$，其具体细节详见表 3.6。为了评估本章方法对这些参数的敏感性，在风景素描数据集上进行实验，每次改变一个超参数，同时保持其他参数为标准值。

表 3.6 中呈现的定量结果表明，当调整这些超参数时，各项性能指标均会受到显著影响。具体而言，与本章方法配置相比，若显著降低任何单个损失分量的权重，均会导致性能大幅下降。这些结果清晰地表明，本章方法的性能对所选取的 λ 值较为

敏感，从而突显了本章方法确定的标准权重对于实现鲁棒图像质量的重要性。

3.2.5 计算开销

表 3.7 计算性能比较。训练指标基于单次迭代进行测量。推理过程的 FLOPs 是生成一个样本所需的计算量。GPU 分配和预留分别指 GPU 的实际分配显存和系统预留显存。所有实验均在 NVIDIA RTX 3090 GPU 上运行，批处理大小为 4。

	训练阶段				推理阶段				
方法	迭代时间 (秒/迭代)	吞吐量 (样本/秒)	GPU 分配 (MB)	GPU 预留 (MB)	吞吐量 (样本/秒)	单样本延迟 (毫秒/样本)	FLOPs (GFLOPs)	GPU 分配 (MB)	GPU 预留 (MB)
StyleGAN2	0.546	7.33	4828.91	7366	138.08	14.27	451.25	953.8	1864
FastGAN	0.392	10.2	2107.63	2644	447.02	7.12	97.36	507.0	770
MixDL	0.711	5.63	7756.45	10674	138.07	14.41	451.25	953.8	1864
PatchDiffusion	0.645	11.8	14458.3	16452	0.22	4462.05	-	1386.3	2538
FastGAN+FAGS	0.444	9.01	3678.71	3828	448.65	7.21	97.36	507.0	770
本章方法	1.221	3.27	21003.23	22126	138.12	14.17	451.25	953.8	1864

本小节在单块 NVIDIA GeForce RTX 3090 GPU 上评估了本章方法与其他对比方法的计算性能。所使用的关键指标包括训练和推理速度、GPU 显存使用量以及浮点运算次数（(Floating Point Operations Per Second，FLOPs)，均汇总于表 3.7。

本章方法的每次训练迭代需要 1.221 秒，并占用 21003 MB 的 GPU 分配显存。尽管表 3.7 显示本章方法的训练成本相较于 FastGAN 等方法更高，但其推理性能表现优异。具体而言，由于 FAGS 和 I&R 模块仅在训练阶段被激活，本章方法的推理指标与基线方法 StyleGAN2 的表现一致，具体体现在其每样本 14.17 毫秒的延迟和 451.25 GFLOPs 的计算量上。

训练阶段较高的开销主要源于 FAGS 模块。在 FAGS 模块内部，若通过对高度为 h、通道数为 c 的特征图采用全局方式来计算测地线自相关一致性损失 L_g，其计算复杂度将近似为 $O(ch^4)$，这将带来过高的计算和显存成本。在章节 3.1.2 和章节 3.2.1 中详述的优化策略，采用基于图像块的方法并结合池化技术，将此特定计算的复杂度显著降低至约 $O(Bch^4/64)$，其中 B 代表批处理大小，从而有效减轻了这一负担。此外，在测地曲面上迭代生成伪源域特征的过程，其具体细节在公式 (3.1) 和 (3.2) 中有所描述。此过程每层的计算复杂度约为 $O(B^2ch^2)$，这也显著增加了整体的训练计算需求和显存占用。因此，在数据受限且不依赖预训练模型的场景中，为了实现从零开始构建伪源域这一优势，这些计算需求构成了一种固有的权衡。在小样本学习的设定下，训练阶段采用这种计算密集型方法，对于充分探索潜在的数据分布并从

固有的少量样本中提取更丰富的语义信息是必要的。

3.2.6 生成图像有效性分析

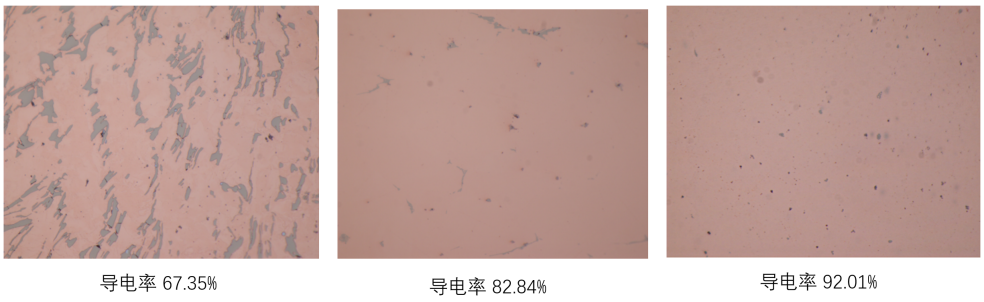


图 3.14　CuCrZr 部分图像示例图

为了验证本章方法在特定领域小样本数据集上的生成效果，以及本章方法在下游任务中的有效性，本节选取了铜铬锆合金（CuCrZr）数据集，用于图像性能预测任务，并以此评估所提出的小样本图像生成方法所生成的图像能否用于数据增强，从而提升模型的泛化能力。

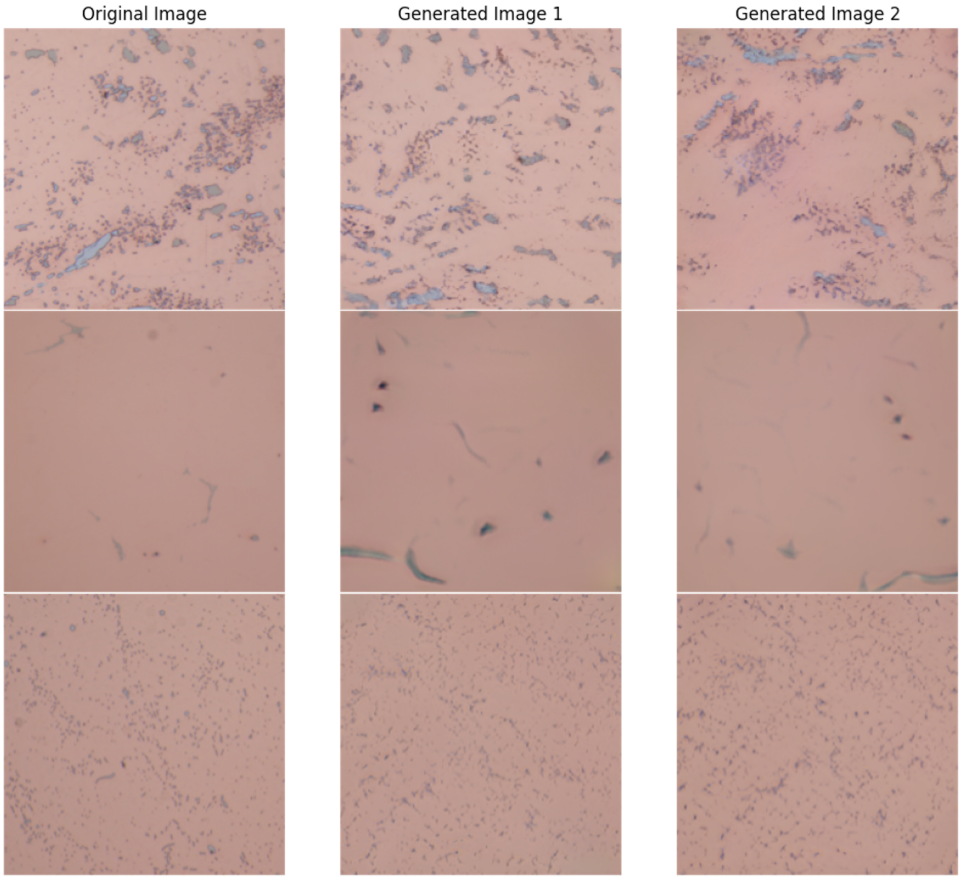


图 3.15　本章方法在 CuCrZr 数据集上的生成图像示例图

对于 CuCrZr 数据集，该数据集包含 18 张微观结构图像，每张图像对应一个不同的导电率标签，部分 CuCrZr 图像的示例如图 3.14 所示。由于样本数量有限且图像尺寸不统一，采用滑动窗口裁剪的方式，将每张原始图像裁剪成 6 - 8 个 224×224 的子图像，从而扩充训练集的规模。每张图像裁剪得到的 6 - 8 个子图像将作为训练集的一部分，每个子图像与其对应的原始图像共享相同的导电率标签。在训练生成模型时，每个对应一张原始图像的标签下的 6 - 8 个子样本被用于训练一个独立的基于本章方法的生成模型，最终共得到 18 个针对特定样本及其标签的生成模型。图 3.15 展示了部分生成图像示例。其中，最左列为裁剪后的原始图像，右侧两列为对应生成模型产生的生成图像样本。

表 3.8 不同特征提取模型的使用与不使用生成式图像增强方法的性能比较

特征提取模型	生成式数据增强	R^2		MSE		MAE		RMSE	
		值	Δ	值	Δ	值	Δ	值	Δ
ResNet18	✗	0.521	-	0.00384	-	0.0490	-	0.0619	-
	✓	0.692	+0.171	0.00247	-0.00137	0.0422	-0.0068	0.0497	-0.0122
ResNet34	✗	0.663	-	0.00270	-	0.0411	-	0.0519	-
	✓	0.728	+0.065	0.00218	-0.00052	0.0405	-0.0006	0.0467	-0.0052
EfficientNet-b5	✗	0.337	-	0.00532	-	0.0606	-	0.0729	-
	✓	0.613	+0.276	0.00310	-0.00222	0.0440	-0.0166	0.0557	-0.0172
DenseNet121	✗	0.485	-	0.00413	-	0.0539	-	0.0643	-
	✓	0.807	+0.322	0.00154	-0.00259	0.0334	-0.0205	0.0393	-0.0250
ConvNeXt	✗	0.448	-	0.00442	-	0.0546	-	0.0665	-
	✓	0.817	+0.369	0.00147	-0.00295	0.0334	-0.0212	0.0383	-0.0282
DeiT	✗	0.611	-	0.00312	-	0.0416	-	0.0558	-
	✓	0.619	+0.008	0.00306	-0.00006	0.0450	+0.0034	0.0553	-0.0005
ViT	✗	0.779	-	0.00177	-	0.0358	-	0.0421	-
	✓	**0.830**	+0.051	**0.00136**	-0.00041	**0.0297**	-0.0061	**0.0369**	-0.0052

针对 CuCrZr 的性能预测任务，使用了多个预训练的图像特征提取模型，包括 ResNet18、ResNet34[94]、EfficientNet-b5[119]、DenseNet121[120]、ConvNeXt[121]、DeiT[122] 以及 ViT[59]，进行图像特征提取。对于 DeiT 和 ViT，选取其输出的 CLS token 作为图像特征，而对于其他模型，选取其全连接层前一层的输出，并经过全局平均池化后作为图像特征。提取得到的特征随后通过 XGBoost[123] 进行回归预测，预测图像特征对应的性能值。实验设置方面，采用了六折交叉验证，在每一折的训练集中，被抽中的 15 张图像，将分别被随机裁剪成 1 张 224×224 的子图像。对于生成式数据增

强，使用 FID 指标来评估为 18 张原始图像分别训练的 18 个生成模型的性能。最终，筛选出其中 FID 值小于 150 的生成模型。在每一折的训练中，对于训练集里的某张原始图像，如果其对应的生成模型属于被筛选出的模型之一，则使用该生成模型生成多张增强图像，连同其原始子图像一起加入该折的训练数据中。每个子图像和生成增强图像都将与原图像共享相同的导电率标签。对于验证集中的 3 张图像，将随机裁剪为 224×224 的图像，不包含任何生成的图像，因此验证集仅包含 3 张原始图像。

表格 3.8 展示了不同特征提取模型的使用与不使用生成式图像增强方法的性能比较，其中 Δ 表示使用生成式增强后相比未使用时的性能变化量。对于 R^2，正 Δ 表示提升，对于 MSE，MAE，RMSE，负 Δ 表示提升。实验结果表明，当生成的图像被加入训练集进行数据增强后，利用任意特征提取模型提取的图像特征训练机器学习模型进行回归预测时，回归指标如 R^2、MSE、MAE 和 RMSE）均有明显提升。例如，使用 DenseNet121 和 ConvNeXt 作为特征提取器时，R^2 值分别从 0.485 和 0.448 提升至 0.807 和 0.817。当选择 ViT 作为特征提取器并使用生成式数据增强时，R^2 指标达到最高值 0.830，其余回归指标也表现出最佳性能。

从图 3.16(a) 和图 3.16(b) 中可以看出，生成图像和原始图像在 t-SNE 空间中的分布。通过观察这三种特征提取模型的可视化结果，生成图像在 t-SNE 空间中的位置与原始图像相对接近，在一些区域又略有扩展，表明生成的图像在特征空间中保持了与原始图像相似的结构，且在一定程度上提升了数据的多样性。

通过比较不同数量的生成图像以及标准增强与生成式增强两种图像增强策略，进一步评估了生成式图像增强在提升模型性能方面的有效性。标准增强包括随机裁剪和水平翻转等常规方法。表 3.9 展示了在不同生成图像数量下，使用标准增强与生成式增强的性能对比结果，这里使用 ViT 作为特征提取模型。其中，生成图像数量指的是每个生成模型生成的图像数量。在每种增强策略下，分别评估了 R^2、MSE、MAE 和 RMSE 等回归指标，并展示了使用生成式增强相比于标准增强的性能变化量 Δ，基线未使用任何增强策略。

从表格中可以看出，在不同数量的生成图像下，生成式数据增强相比于标准增强普遍带来了性能的提升。特别是在生成图像数量为 4 和 10 时，生成式数据增强显著提高了模型性能，并且超越了未使用增强的情况。在生成 4 张图像进行增强时，模型性能提升最大，R^2 从 0.552 提升至 0.830，MSE 从 0.00359 降至 0.00136，MAE 从

表 3.9　不同增强图像数量与图像增强策略的性能比较

增强图像数量	图像增强策略	R^2		MSE		MAE		RMSE	
		值	Δ	值	Δ	值	Δ	值	Δ
0	无	0.779	-	0.00177	-	0.0358	-	0.0421	-
2	标准	0.678	-	0.00258	-	0.0445	-	0.0508	-
	生成式	0.667	-0.011	0.00267	+0.00009	0.0399	-0.0046	0.0516	+0.0008
4	标准	0.552	-	0.00359	-	0.0472	-	0.0599	-
	生成式	**0.830**	+0.278	**0.00136**	-0.00223	**0.0297**	-0.0175	**0.0369**	-0.0230
8	标准	0.528	-	0.00378	-	0.0510	-	0.0615	-
	生成式	0.537	+0.009	0.00371	-0.00006	0.0499	-0.0011	0.0609	-0.0006
10	标准	0.606	-	0.00316	-	0.0463	-	0.0562	-
	生成式	0.798	+0.192	0.00162	-0.00154	0.0327	-0.0136	0.0402	-0.0160
16	标准	0.289	-	0.00570	-	0.0649	-	0.0755	-
	生成式	0.428	+0.139	0.00458	-0.00112	0.0568	-0.0081	0.0677	-0.0078

0.0472 降至 0.0297，RMSE 从 0.0599 降至 0.0369。这些结果表明，生成图像能够有效提升数据多样性，并改善模型的泛化能力，且在该设置下模型性能达到了最佳。然而，当生成图像数量过多或过少时，反而可能影响模型性能，导致其相较于未使用增强策略的性能有所下降。

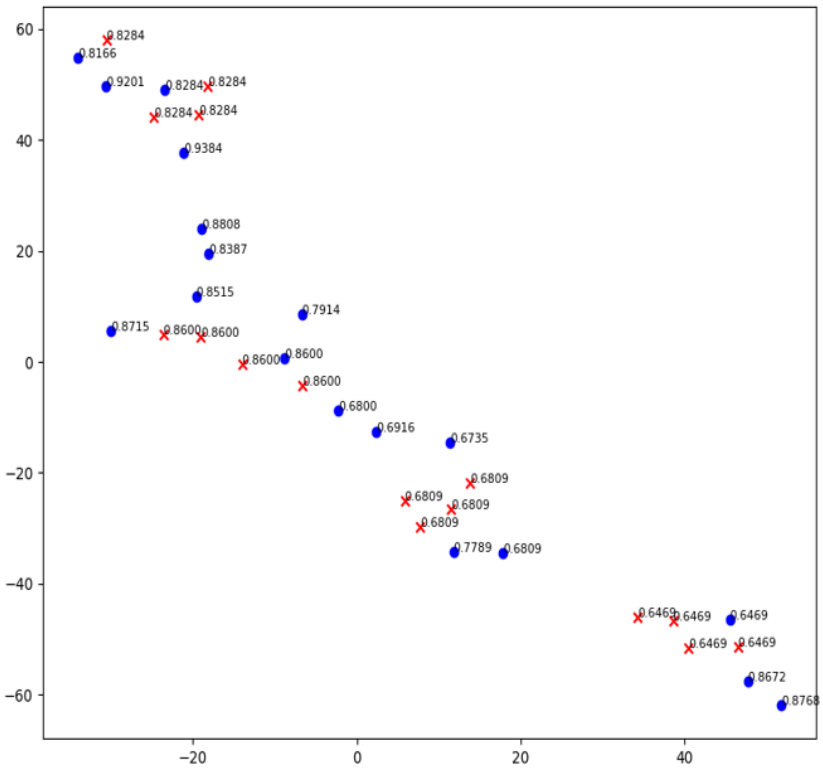
从图 3.17 和图 3.18 中的 t-SNE 可视化图可以看出，随着生成图像数量的增加，生成式数据增强对特征空间的影响逐渐增大。图中蓝色圆点表示原始图像，红色叉号表示生成图像。如图 3.17(a) 所示，在标准增强的情况下，生成图像的分布相对集中，且与原始图像的位置接近，说明标准增强方法增加的多样性较为有限，生成图像的特征分布未能显著拓展数据空间。然而，当使用生成式增强时，生成图像的分布逐渐扩展，如图 3.17(b) 所示。在生成 4 张图像进行增强时，如图 3.18(a)，生成图像的分布明显扩展，与原始图像有了更大的区分，且生成图像的多样性提高，R^2、MSE、MAE 和 RMSE 等性能指标在表 3.9 中也显示出显著的提升。在生成图像数量为 10 张时，见图 3.18(b)，生成图像的分布仍较为分散，可能导致数据增强的效果受到一定限制，但它们在特征空间中的位置保持了一定的聚集趋势，这有助于增强模型的泛化能力。

总体来看，t-SNE 可视化结果与性能指标共同表明，本章提出的生成式增强方法能够比标准增强更有效地扩展数据的特征空间分布，提升数据多样性。在增强图像数量适中时，这种多样性的提升能够显著改善下游任务模型的性能和泛化能力。然

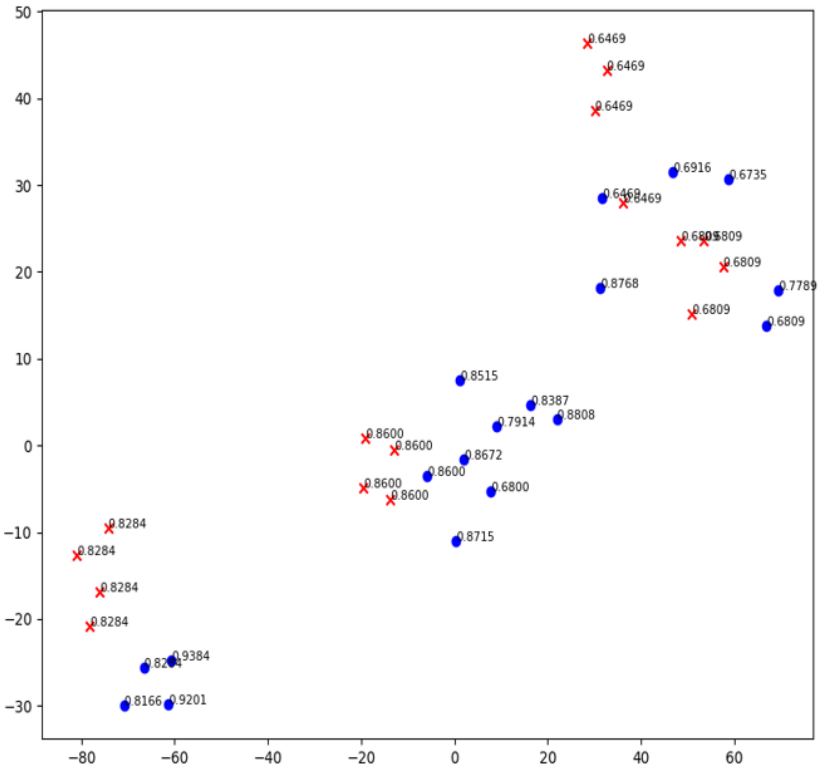
而，当生成图像数量过多时，尽管特征空间覆盖范围更广，但性能反而下降，甚至低于基线。这可能表明，过多的生成样本中可能包含了一些偏离原始数据流形的样本，反而对模型训练产生了干扰。这种现象或许与本章方法主要关注于生成与训练样本相似的图像，对生成样本的多样性缺乏显式、精细的控制有关，这是未来值得改进的方向。

3.3 本章小结

本章提出了一种基于预形状空间中测地曲面信息迁移的小样本图像生成方法，旨在解决特定专业领域中因数据采集困难而导致的小样本问题。通过在小样本场景中采用特征增强策略，结合测地曲面特征增强模块与插值监督和正则化模块，有效克服了传统生成模型在小样本环境中的瓶颈。在此框架下，训练得到的生成器能够生成高保真度且具有多样性的图像，从而有效扩充数据集，进而提升模型在小样本学习任务中的表现。实验结果表明，所提出的方法在多种类的小样本数据集上都展现出了较好的性能，并且在小样本材料数据集上也可生成有效图像，提升下游任务性能，具有一定的应用价值。

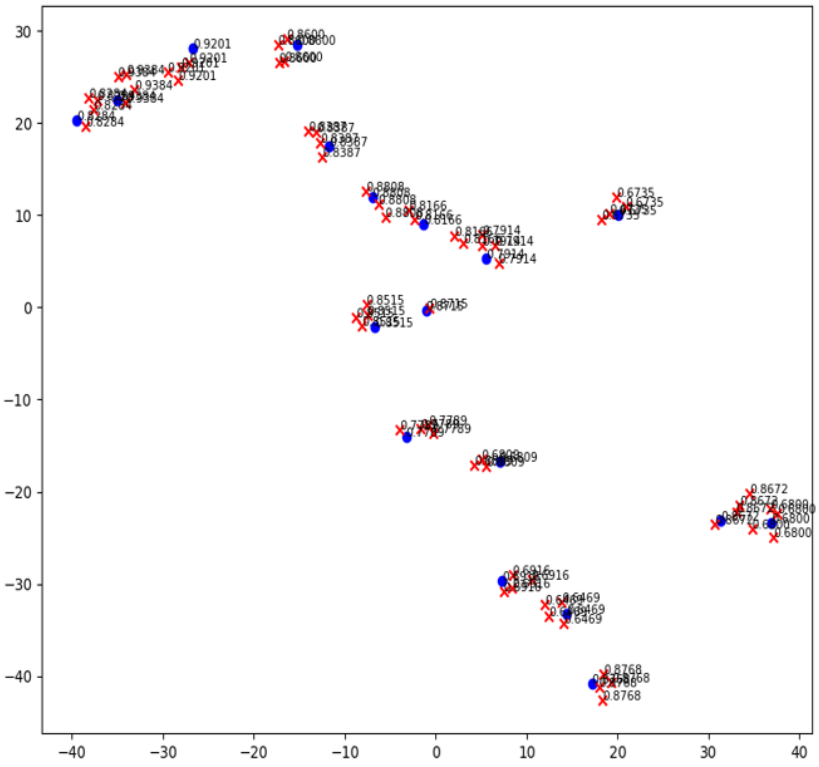


(a)

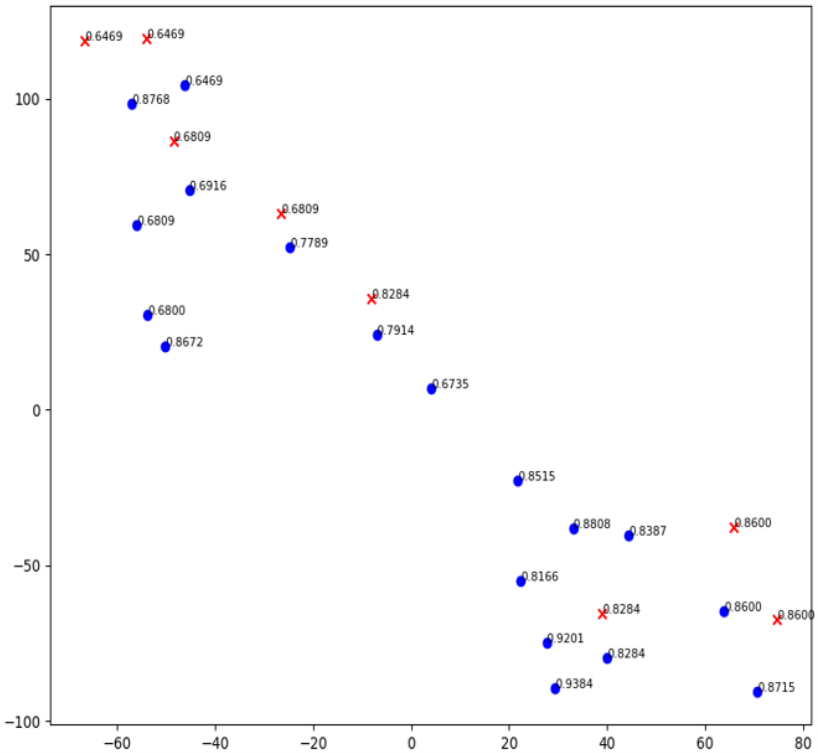


(b)

图 3.16　不同特征提取模型提取图像特征分布情况的 t-SNE 可视化图：(a) DeiT 提取特征的 t-SNE 可视化图；(b) DenseNet-121 提取特征的 t-SNE 可视化图。图中蓝色圆点表示原始图像，红色叉号表示生成图像。

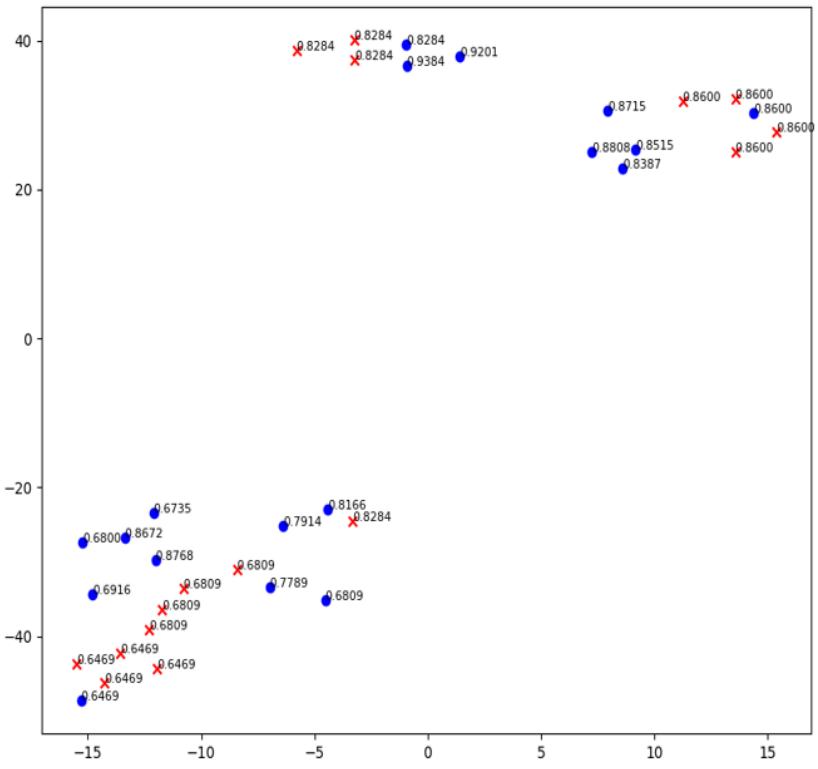


(a)

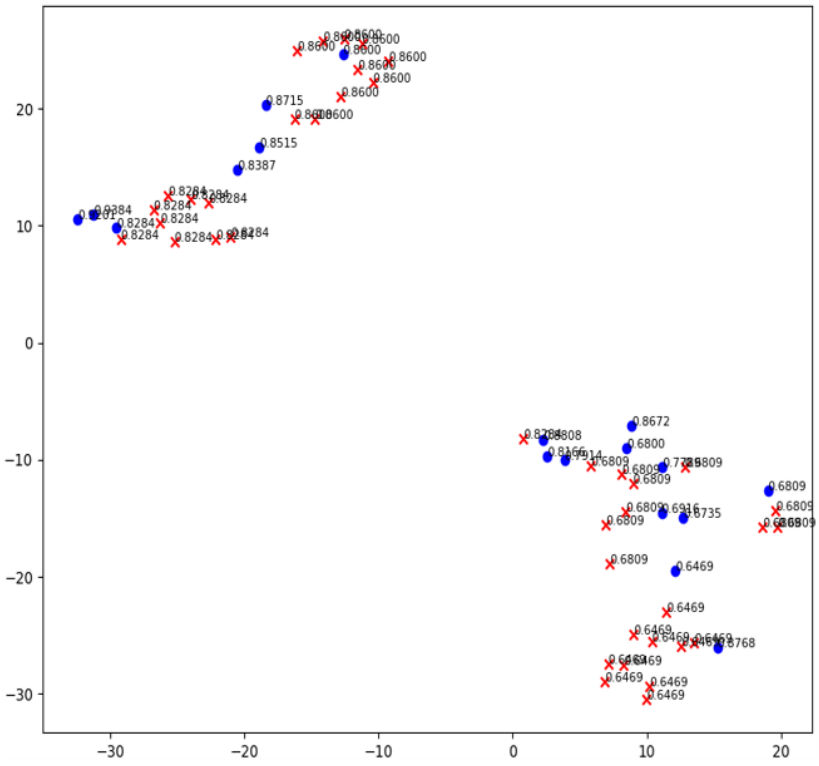


(b)

图 3.17　不同生成图像数量下，标准增强与生成式增强的 t-SNE 可视化图，其中生成图像数量指的是每个生成模型生成的图像数量。(a) 使用标准增强的 t-SNE 可视化图；(b) 使用生成式增强的 t-SNE 可视化图，生成图像数量为 2。图中蓝色圆点表示原始图像，红色叉号表示生成图像。



(a)



(b)

图 3.18　不同生成图像数量下，标准增强与生成式增强的 t-SNE 可视化图，其中生成图像数量指的是每个生成模型生成的图像数量。(a) 使用生成式增强的 t-SNE 可视化图，生成图像数量为 4；(b) 使用生成式增强的 t-SNE 可视化图，生成图像数量为 10。图中蓝色圆点表示原始图像，红色叉号表示生成图像。

第四章　基于预形状空间中测地曲面增强的零样本文本引导图像风格迁移

继第三章验证了测地曲面特征增强策略在小样本生成中的有效性后，本章将其应用于零样本文本引导的图像风格迁移。不同于上一章从零训练生成对抗网络，本章运用此增强思想指导预训练扩散模型的推理过程，使其作用从数据增强转变为促进风格与内容的融合，且无需额外训练。基于此，本章提出一种新的风格迁移方法，其核心是将测地曲面特征增强机制创新性地融入扩散模型推理引导，结合预形状自相关一致性约束，旨在克服传统风格迁移方法对参考图像的依赖以及在内容保持上的不足，实现高效、灵活、高质量且内容保持良好的文本控风格生成。

4.1　方法概述

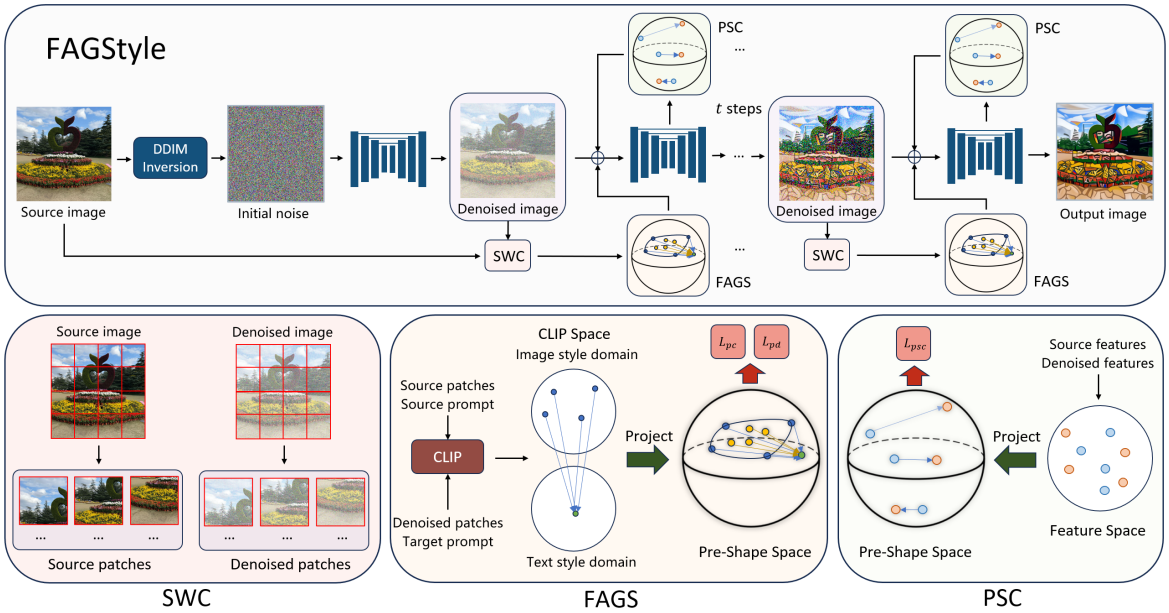


图 4.1　基于 FAGS 的零样本文本引导图像风格迁移方法的流程图。该方法旨在更精准地引导扩散模型进行风格迁移。其核心在于，通过融合滑动窗口裁剪、测地曲面特征增强以及预形状自相关一致性三个模块来分别改进风格和内容控制损失。在模型的推理阶段，这些优化后损失函数的梯度会在每个去噪时间步被施加到去噪图像上，从而实现对风格迁移过程的精确引导。

本章基于预形状空间中测地曲面增强（Feature Augmentation on Geodesic Surface，

FAGS）的零样本文本引导图像风格迁移方法（FAGStyle）可以分为两个部分：对于参考文本风格信息的迁移以及对于源图中内容信息的保持。图 4.1 展示了本章方法的流程。首先，本章仅采用一个在真实照片图像数据集上预训练的扩散模型（Diffusion Model，DM），输入风格参考文本，提取风格文本特征后，将其融合进 DM 的推理过程中，以此完成零样本文本引导的风格迁移。具体来说，是使用生成图像、源图像、风格参考文本计算风格控制损失和内容控制损失，再将以上损失的梯度加入到扩散模型推理过程中各个时间步得到的去噪图像上。其中，对于风格控制损失的计算，提出滑动窗口图像块裁剪（Sliding Window Crop，SWC），再使用 FAGS 融合各图像块信息，增强信息交互。对于内容控制损失，提出预形状自相关一致性（Pre-Shape Self-correlation Consistency，PSC）模块，确保风格化图像与源图内容的结构信息一致。本小节将对基于文本信息引导的扩散模型风格迁移范式，基于 FAGS 模块特征融合的风格控制损失计算以及预形状自相关一致性模块进行介绍。

4.1.1 基于文本信息引导的扩散模型风格迁移

如章节 2.2.2 所述，扩散模型通过学习反向去噪过程生成图像，其核心是噪声预测网络 $\epsilon_\theta(x_t, t)$。去噪扩散隐式模型（Denoising Diffusion Implicit Model，DDIM）[124] 对此过程提供了一个更通用的采样框架，其单步更新公式形式如下：

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{x}_{0,t}(x_t) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t, t) + \sigma_t\,\epsilon, \tag{4.1}$$

此处，$\hat{x}_{0,t}(x_t)$ 代表基于 x_t 估计出的去噪图像，其计算方式为：

$$\hat{x}_{0,t}(x_t) := \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{4.2}$$

其中噪声调度参数 $\alpha_t, \bar{\alpha}_t$ 及噪声预测网络 ϵ_θ 的定义与第 2.2.2 节一致，$\epsilon \sim \mathcal{N}(0, I)$ 为标准高斯噪声。参数 σ_t 控制采样的随机性。虽然设置 $\sigma_t = 0$ 可以实现确定性 DDIM 采样，有利于最大化内容保持，但有时可能会限制生成结果的多样性和风格化的强度。为了获得更强的风格化效果，本章方法选择保留与 DDPM[3] 一致的随机性，这通过将 σ_t 按下式设置来实现：

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{4.3}$$

其中 $\beta_t = 1 - \alpha_t$。这种选择旨在利用随机性增强生成结果的风格表现力。

在此随机采样框架下，为了在采样过程中施加风格引导，借鉴了分类器引导（Classifier Guidance）[104] 机制。在每个去噪时间步 t，首先依据公式 (4.2) 计算出当前估计的去噪图像 $\hat{x}_{0,t}(x_t)$。然后，利用一个外部定义的总损失函数 L_{total} 计算其对该去噪图像的梯度 $\nabla_x L_{\text{total}}$。该梯度随后被用于调整 $\hat{x}_{0,t}(x_t)$，得到一个引导后的去噪图像 $\hat{x}'_{0,t}(x_t)$，使其更符合目标风格：

$$\hat{x}_{0,t}(x_t) = \hat{x}_{0,t}(x_t) + \nabla_x L_{\text{total}}(x)\Big|_{x=\hat{x}_{0,t}(x_t)}. \tag{4.4}$$

这个经过引导的去噪图像估计 $\hat{x}_{0,t}(x_t)$ 将替代公式 (4.2) 中原始的 $\hat{x}_{0,t}(x_t)$，用于计算得到更符合目标风格的上一步图像 x_{t-1}。总损失 L_{total} 通常结合了风格损失 L_{sty} 与内容损失 L_{cont}：

$$L_{\text{total}} = \lambda_{sty} L_{\text{sty}} + \lambda_{cont} L_{\text{cont}}, \tag{4.5}$$

其中 λ_{sty} 和 λ_{cont} 是平衡权重。关于 L_{sty} 和 L_{cont} 的具体设计将在后续小节中详细阐述。

4.1.2 基于测地曲面特征增强模块的特征融合

由于 CLIP 在多模态特征提取方面的强大能力，大多数现有的风格控制损失都基于由 CLIP 提取得到的特征进行计算，例如全局 CLIP 损失[12] 和方向 CLIP 损失[69]。以往的研究通常是将整张图像输入 CLIP ，进行特征提取，用于风格控制损失的计算。CLIPStyler[125] 通过从图像中随机裁剪出多个图像块（Patch）并提取其 CLIP 特征，实现了风格迁移效果的提升。利用这些基于局部块的 CLIP 特征，模型可以更聚焦于纹理、色彩分布等局部风格信息，减少了全局特征对主体内容的过度依赖，并有助于捕捉具有空间不变性的风格语义。

然而，随机裁剪的图像块可能会过于集中在某些区域，从而导致这些区域被过度风格化，而其他区域则风格化不足。此外，随机裁剪可能会破坏源图像中内容的主要结构，如消融实验中图 4.11 第 1 行所示，且裁剪出的图像块之间缺乏空间关联性。为解决上述问题，提出了滑动窗口裁剪（Sliding Window Crop，SWC）方法。该方法通过设定一个固定大小的窗口和小于窗口尺寸的固定步幅（Stride），让窗口在图像上按预定路径滑动并依次提取图像块。这种方式裁剪出的图像块具有相互重叠部分，并能完整覆盖整张图像，不仅使后续提取的 CLIP 特征在空间上具有连续性约

束，也保留了基于图像块计算损失所能提供的空间不变性信息，有助于迫使生成器在所有空间位置都满足风格一致性。

在时间步 t 时，假设去噪后的图像 $\hat{x}_{0,t}(x_t)$ 和源图像 x_0 都被缩放到 $H \times W$ 大小，其中 $H = W$。将图像块的数量设为 n，作为一个超参数，并把图像通过滑动窗口的形式均匀地裁剪成 $n_w = n_h = \sqrt{n}$ 个正方形的图像块。每个裁剪得到的图像块的高度和宽度为 $H_p = W_p = W/(n_w + 1) = H/(n_h + 1)$。滑动窗口的步幅（Stride）设为 $s = H_p/2 = W_p/2$，使得每个图像块与其相邻的图像块之间在边界上有一半部分的重叠，使得图像块之间保有一定空间关联性，实现图像块之间信息的交互。对图像 $\hat{x}_{0,t}$ 和 x_0 而言，通过 SWC 得到的第 i 个图像块 $\hat{x}^i_{0,t}$ 和 x^i_0 的左上角坐标 (e, f) 定义如下：

$$(e, f) = \left(\left\lfloor \frac{i}{n_h} \right\rfloor \cdot s,\ (i \bmod n_w) \cdot s \right), \tag{4.6}$$

其中 $i = 0, 1, \ldots, n-1$，$\lfloor \cdot \rfloor$ 表示向下取整，mod 表示取模运算。

所提出的 SWC 方法会根据输入图像的大小和需要裁剪的图像块数动态地调整窗口大小和步幅。这种方式确保了相邻图像块的重叠部分，实现了有效的信息交互，使得整张图像的风格迁移更均匀、更一致。

虽然 SWC 能够在相邻图像块之间实现一定的信息交互，并融合了一定位置信息，但仍须解决非相邻图像块间缺乏信息交互的问题。当图像块数量较多时，第一个与最后一个图像块之间可能存在较大的空间距离，缺乏交互会导致不同区域的过度或不足风格化，从而影响生成图像的整体质量。

为了解决上述问题，引入了第三章中提出的测地曲面上特征增强（Feature Augmentation on Geodesic Surface，FAGS）策略。根据 FAGS 策略，先将通过 SWC 得到的所有图像块的特征投影到预形状空间中，以构建一个测地曲面。测地曲面上的任一点都可视为增强后的新特征，该新特征中带有来自所有裁剪得到的图像块的加权信息，包含全局上下文信息。然后，根据所构建测地曲面上的这些增强特征来计算风格控制损失。

具体来说，利用预训练 CLIP 模型的图像编码器 E_{img}，从去噪后的图像裁剪得到的图像块提取特征 $\bar{x}_t$:

$$\bar{x}_t = \left\{ E_{\text{img}}\left(\hat{x}^i_{0,t}\right) \;\middle|\; E_{\text{img}}\left(\hat{x}^i_{0,t}\right) \in \mathbb{R}^{c \times h \times w},\ i = 1, 2, \ldots, n \right\}, \tag{4.7}$$

以及从源图像裁剪得到的图像块中提取特征 $\bar{x}_0$:

$$\bar{x}_0 = \Big\{ E_{\text{img}}\big(x_0^i\big) \;\Big|\; E_{\text{img}}\big(x_0^i\big) \in \mathbb{R}^{c\times h\times w},\, i = 1, 2, \ldots, n \Big\}. \tag{4.8}$$

为构建测地曲面,这些特征必须首先被重塑(Reshape)并投影到预形状空间(Pre-Shape Space)中。重塑操作 $\mathcal{R}(\cdot)$ 将每个维度为 $c \times h \times w$ 的特征张量转换为维度为 $2 \times (chw/2)$ 的形式，将其解释为二维空间中的地标点。随后，使用均值消减和归一化的组合操作 $f_p(\cdot) = \mathcal{V}\big(\mathcal{Q}\big(\mathcal{R}(\cdot)\big)\big)$ 将特征投影到预形状空间中。

在预形状空间中，可以构建测地曲面 $\mathbb{G}_{surf}$[23]，其定义为:

$$\mathbb{G}_{surf}\big(\tau, \omega\big) = \mu_n, \tag{4.9}$$

其中 μ 同时是一个向量，也是测地曲面 $\mathbb{G}_{surf}$ 上的一个预形状，又下式计算得到:

$$\mu_j = \mathbb{G}_{cur}\big(\mu_{j-1}, \tau_j\big)\Big(\frac{\omega_j}{\sum_{i=1}^{j}\omega_i}\Big), \quad \text{其中} j = 2, \ldots, n, \tag{4.10}$$

$\tau \triangleq [\tau_1, \ldots, \tau_n]^T$ 和 $\omega \triangleq [\omega_1, \ldots, \omega_n]^T$ 分别代表给定的一组向量和对应的权重，ω 控制着图像块特征之间信息交互的程度。n 表示向量个数，在本场景中与裁剪得到的图像块数量一致。$\mu_1 = \tau_1$，因此当 $j = n$ 时，可以使用一组向量 τ 和一组权重 ω 构建一个测地线表面 $\mathbb{G}_{surf}$。具体来说，本章方法以去噪图像特征 $f_p(\bar{x}_t)$ 和原始图像特征 $f_p(\bar{x}_0)$ 作为输入，构建了两个测地曲面。

通过输入不同的 m 组权重，便可在各自测地曲面上得到多组增强特征向量 $\tilde{x}_t^i$ 和 $\tilde{x}_0^i$，此处 $i = 1, 2, \ldots, m$。此处，令 $m = n$，即保证增强特征数与裁剪得到的图像块数相同。每个增强特征向量都整合了来自所有图像块的全面交互信息，增强了迁移过程中图像的信息一致性和风格准确性。

本章方法在原有的基础上，将 SWC 与在预形状空间上进行的 FAGS 策略相结合，提出了风格控制损失函数的改进版本，其形式如下:

$$L_{\text{sty}} = \lambda_{pc}\, L_{pc}\big(\hat{x}_{0,t}, p_{\text{tgt}}\big) + \lambda_{pd}\, L_{pd}\big(\hat{x}_{0,t}, x_0, p_{\text{tgt}}, p_{\text{src}}\big), \tag{4.11}$$

其中 L_{pc} 和 L_{pd} 分别在 SWC 的基础上结合了 FAGS 策略，旨在改进已有的图像块 CLIP 损失（PatchCLIP Loss）[125] 和图像块方向损失（Patch Directional Loss）[73]。此处，p_{tgt} 和 p_{src} 分别表示用于参考的目标风格文本描述（Target Prompt）以及源图像

风格的文本描述（Source Prompt）。如果输入图像是真实照片，则将 p_{src} 设为“照片（Photo）”。λ_{pc} 和 λ_{pd} 分别是对应 L_{pc} 和 L_{pd} 的超参数权重。

L_{pc} 项通过计算构建出的测地曲面上增强后的源图像的图像块特征 $\tilde{x}_t^i$ 与由文本提示 p_{tgt} 得到的目标 CLIP 特征之间的测地距离来实现风格对齐，具体如下式所示：

$$L_{pc}\big(\hat{x}_{0,t}, p_{\text{tgt}}\big) = \frac{1}{n}\sum_{i=1}^{n} d\big(\tilde{x}_t^i,\ f_p\big(E_{\text{txt}}\big(p_{\text{tgt}}\big)\big)\big), \tag{4.12}$$

其中 E_{txt} 表示 CLIP 的文本编码器。由于 $\tilde{x}_t^i$ 位于预形状空间中，也需将目标 CLIP 文本特征 $E_{\text{txt}}(p_{\text{tgt}})$ 投影至预形状空间，才能计算它们之间的测地距离 $d(\cdot,\cdot)$。

L_{pd} 项通过对齐源图像——文本风格特征与目标图像-文本风格特征在预形状空间中的方向来实现风格一致性，公式如下：

$$L_{pd}\big(\hat{x}_{0,t}, x_0, p_{\text{tgt}}, p_{\text{src}}\big) = \frac{1}{n}\sum_{i=1}^{n}\Big(1 - \frac{\langle \Delta I_i, \Delta T\rangle}{\|\Delta I_i\| \cdot \|\Delta T\|}\Big), \tag{4.13}$$

其中使用从测地曲面上获得的增强源图像的图像块特征 $\tilde{x}_0^i$ 和增强目标图像的图像块特征 $\tilde{x}_t^i$，定义它们在预形状空间中的方向向量：

$$\Delta I_i = \tilde{x}_t^i - \tilde{x}_0^i. \tag{4.14}$$

类似地，将源文本特征 $E_{\text{txt}}(p_{\text{src}})$ 和目标文本特征 $E_{\text{txt}}(p_{\text{tgt}})$ 也投影到预形状空间中，从而得到它们在预形状空间中的方向：

$$\Delta T = f_p\big(E_{\text{txt}}\big(p_{\text{tgt}}\big)\big) - f_p\big(E_{\text{txt}}\big(p_{\text{src}}\big)\big). \tag{4.15}$$

通过此改进的风格控制损失，在提供局部图像块中空间不变性信息的同时，保证了相邻与非相邻图像块之间的信息交互，局部与全局语义信息共同作用，从而在整张图像上实现风格迁移的均匀与一致性。此外，风格控制损失中关于图像特征和文本特征之间的计算均在预形状空间中进行，相比在 CLIP 图像空间与 CLIP 文本空间两个不同的空间中，投影至同一预形状空间使得特征之间的距离被更加准确的度量所表示。

4.1.3 预形状自相关一致性模块

在风格迁移的过程中同时保证有效的风格化以及内容的保持是非常具有挑战性的，虽然上一节改进的风格控制损失可以增强风格迁移过程中风格信息的表达，使

得图像被更均匀一致地风格化，但在某些情况下可能会削弱内容保留能力。因此，在本小节中，在内容控制损失中引入预形状自相关一致性（Pre-Shape Self-correlation Consistency，PSC）模块，进一步确保在风格迁移过程中的风格化和内容保持的平衡。

在上一章中，提出了测地自相关一致性损失（Geodesic Self-correlation Consistency Loss），将预形状空间中的目标图像特征与在预形状空间中测地曲面上生成的伪源域图像特征分别计算各自的自相关矩阵，确保生成目标图像与增强后源图像的内在结构一致性。本节中所考虑的风格迁移任务，也需要考虑到生成图像对源图内容的保持，因此，引入了预形状自相关一致性损失（Pre-Shape Self-correlation Consistency Loss，L_{psc}）以确保生成图像中内容与源图内容的内在结构一致。

Baranchuk 等人[126] 的最新研究表明，扩散模型中 U-Net 噪声预测器 ϵ_θ 对输入图像提取得到的中间特征保留了图像的空间信息。由于自相关一致性损失在特征层面进行自相关矩阵的计算，将 ϵ_θ 用作特征提取器。在扩散模型反向去噪过程的每一步中，源图像 x_0 和在当前步 t 的去噪后图像 $\hat{x}_{0,t}$ 都会被输入到 ϵ_θ。利用 ϵ_θ 的编码器部分 E_{ϵ_θ} 分别提取出 x_0 和 $\hat{x}_{0,t}$ 的特征图。

在上一章的小样本图像生成的任务设置中，数据集中存在多张源图像，提出的方法通过 FAGS 策略对多张源图像提取得到的特征进行增强，用于计算测地自相关一致性损失。然而，在本章设置的场景中仅有一张源图像 x_0 与一张目标图像 $\hat{x}_{0,t}$。由于将图像裁剪为图像块可能会破坏图像中内容的完整性性，因此在内容控制损失的计算中，并未将图像拆分成多个图像块并使用 FAGS 进行特征增强，而是直接将完整的源图像与目标图像输入 E_{ϵ_θ} 提取得到的特征投影到预形状空间，再进行后续的计算，具体如下：

$$z_0^l = f_p\big(E_{\epsilon_\theta}(x_0)\big) \in \mathbb{R}^{c\times h\times w}, \tag{4.16}$$

以及

$$\hat{z}_{0,t}^l = f_p\big(E_{\epsilon_\theta}\big(\hat{x}_{0,t}\big)\big) \in \mathbb{R}^{c\times h\times w}. \tag{4.17}$$

在得到位于预形状空间中的源图像特征 z_0^l 和目标图像特征 $\hat{z}_{0,t}^l$ 后，计算它们各自的自相关矩阵。设 $z_0^l(u)$ 与 $z_0^l(v)$ 分别表示源图像特征在位置 $u, v \in \mathbb{R}^{h\times w}$ 处的 c

维向量，则下式给出了源图像特征中位于 u 与 v 位置上的向量之间的点积：

$$C_{u,v}^{z_0^l} = \langle z_0^l(u), z_0^l(v) \rangle, \tag{4.18}$$

遍历特征图中在 $h \times w$ 维度上的所有位置，可得到源图像特征 z_0^l 的自相关矩阵 $C^{z_0^l}$。同理，可得目标图像特征 $\hat{z}_{0,t}^l$ 的自相关矩阵 $C^{\hat{z}_{0,t}^l}$。预形状自相关一致性损失 L_{psc} 定义如下：

$$L_{psc}\big(x_0, \hat{x}_{0,t}\big) = \mathbb{E}_{x_0} \sum_l L_{s\ell 1}\Big(C^{z_0^l}, C^{\hat{z}_{0,t}^l}\Big), \tag{4.19}$$

其中 l 遍历噪声预测器中编码器部分的 E_{ϵ_θ} 的若干选定层，(u, v) 遍历特征空间中的所有位置，$L_{s\ell_1}(\cdot)$ 表示 Smooth - ℓ_1 损失函数[110]。通过该损失项，能够在风格迁移的过程中更好地保留源图像中内容的内在结构信息，从而使得目标图像中内容的结构、形状等信息保持不变。

加入 L_{psc} 后的内容控制损失 L_{cont} 定义如下：

$$L_{\text{cont}} = \lambda_{ps}\, L_{psc}\big(x_0, \hat{x}_{0,t}\big) + \lambda_z\, L_{ZeCon}\big(x_0, \hat{x}_{0,t}\big) + \lambda_v\, L_{VGG}\big(x_0, \hat{x}_{0,t}\big) + \lambda_m\, L_{MSE}\big(x_0, \hat{x}_{0,t}\big). \tag{4.20}$$

其中，额外引入的 L_{ZeCon} 用于计算从 x_0 和 $\hat{x}_{0,t}$ 提取的特征之间的交叉熵损失[73]。L_{VGG} 则通过最小化两张图像的 VGG 特征图之间的均方误差来保留源图中的内容，L_{MSE} 用 ℓ_2 范数来衡量两张图像在像素层面的差异。各损失函数所对应的权重 λ 皆为固定的超参数。在后续消融实验小节中，给出具体的推荐超参数设置。

4.2 实验分析

本节首先介绍实验设置和数据集，包括多种源图像和多种不同的参考风格文本，接着，展示在多种源图像和风格文本上的实验结果，包括在想象类风格以及常见类风格上的定性对比实验、定量对比实验以及各个模块的消融实验和超参数设置实验。

4.2.1 实验设置

本节所有实验均在相同的超参数设置下进行，具体使用的超参数可以参见后续消融实验小节 4.2.5 的超参数设置部分。对于本章方法使用的基础生成模型，选用在 ImageNet 数据集上预训练的无条件扩散模型 (Unconditional Diffusion Model，

图 4.2 所用数据集中 25 张输入源图像的实例

UDM)[104] 。选择 UDM 的目的是为了确保对本章方法进行严格的零样本 (zero-shot) 评估。这一点至关重要，因为 UDM 主要在 ImageNet 这类包含大量无特定艺术风格的照片数据集上进行训练。因此，可以认为该预训练模型本身几乎不具备关于多样化艺术风格的先验知识，这与许多在包含丰富风格信息的网络数据，例如 LAION 数据集，上训练的现代扩散模型，诸如 Stable Diffusion[127] 或 Flux[128] 等，形成了鲜明对比。此类模型在预训练阶段可能已经学习并记忆了大量的风格模式。因此，通过使用一个风格先验较少的 UDM 作为基础模型，可以最大限度地减少来自预训练知识的混淆因素，从而能够更清晰、更可靠地验证 FAGStyle 方法仅根据文本输入引导生成模型此前未曾接触过的新颖风格的能力。此外，UDM 简单的架构便于集成 FAGS

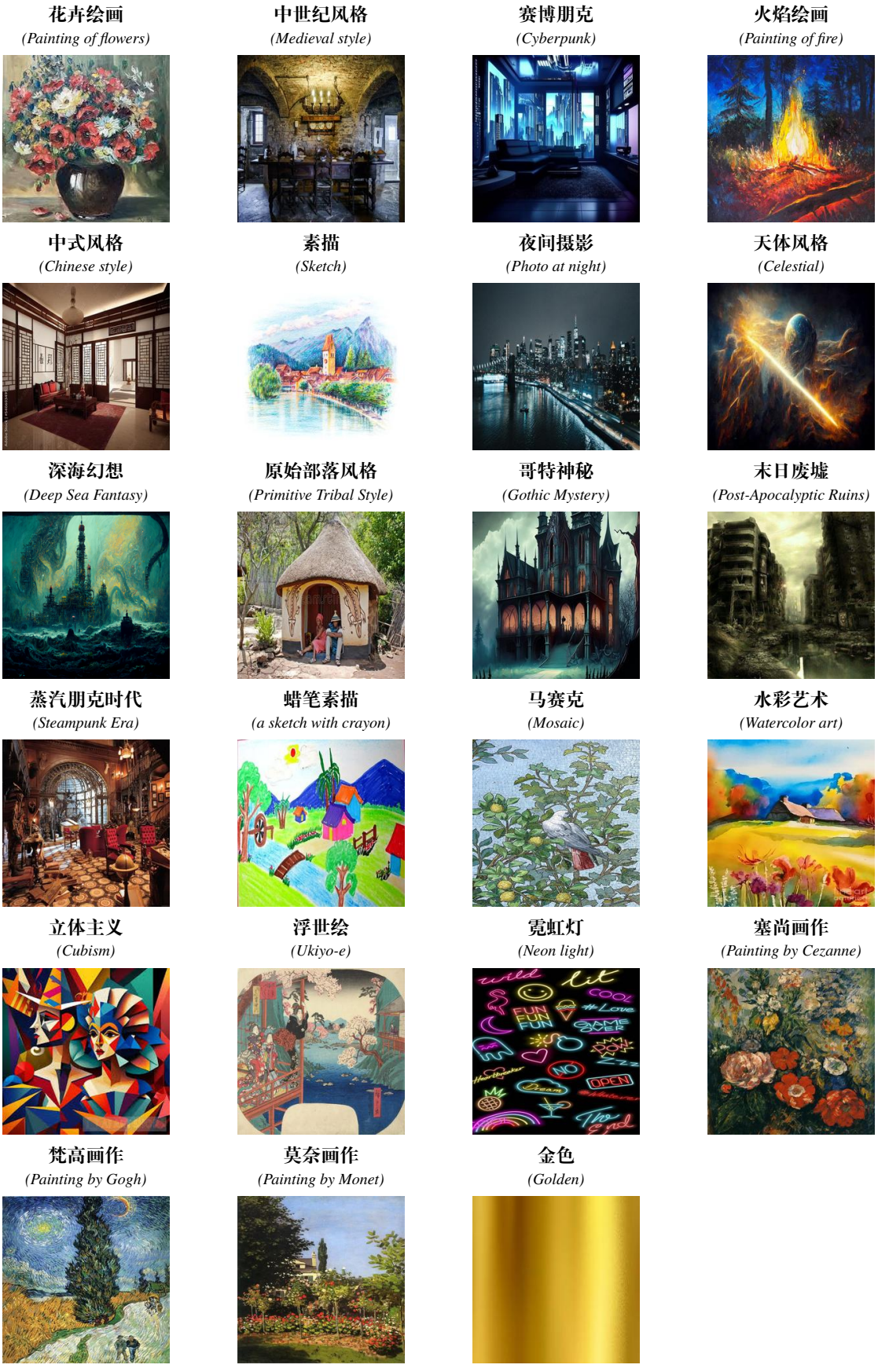


图 4.3 所用数据集中 23 个常见类风格的文本提示以及参考图像

表 4.1　所用数据集中 32 个想象类风格的文本提示

中文翻译	英文提示词
混沌绘画	Painting of chaos
赛博朋克混搭立体主义	Cyberpunk mixing Cubism
火与花的画作	Painting of fire and flowers
虚无画作	Painting of void
中式哥特神秘风格	Chinese Gothic Mystery
科幻小说插画	Sicence Fiction Illustration
中世纪哥特神秘风格	Medieval Gothic Mystery
蒸汽朋克幻想	Steampunk Fantasy
赛博朋克混搭中国风	Cyberpunk mixing Chinese style
深海中世纪风格	Deep Sea Medieval Style
冰封原始部落风格	Frozen Primitive Tribal style
中国复古电影风格	Chinese Vintage film
赛博朋克素描	Cyberpunk Sketch
超现实原始部落风格	Surreal Primitive Tribal Style
蒸汽朋克废墟	Steampunk Ruins
深海原始部落风格	Deep Sea Primitive Tribal Style
深海哥特神秘风格	Deep Sea Gothic Mystery
丛林蒸汽朋克	Jungle Steampunk
深海蒸汽朋克	Deep Sea Steampunk
深海赛博朋克	Deep Sea Cyberpunk
丛林赛博朋克	Jungle Cyberpunk
丛林幻想	Jungle Fantasy
中世纪丛林幻想	Medieval Jungle Fantasy
超现实中世纪风格	Surreal Medieval Style
中世纪幻想	Medieval Fantasy
冰封丛林	Frozen Jungle
深海下的宇宙	Universe under Deep Sea
幻想	Fantasy
梦境画作	Painting of a dream
宇宙	Universe
超现实梦境	Surreal Dreamscape
末日废墟	Post-Apocalyptic Ruins

模块。本次实验中所使用的 UDM 模型，其训练采用的图像分辨率为 256×256。对于图像和文本中风格信息的提取，使用了预训练的 CLIP[11] 作为图像和文本编码器进行。在前向扩散阶段，进行 DDIM 反演，对输入源图像进行加噪处理，并将默认的总采样时间步从 $T = 1000$ 调整为 $T' = 50$[73]，同时设置初始噪声时间步 $t_0 = 25$ 。

在反向采样过程中，采用 DDPM 策略来获得更多样的风格化结果。本章方法在严格的零样本设置下运行，利用前述的、仅在无风格的真实世界图像上预训练的 UDM。值得注意的是，预训练 UDM 在推理过程中直接用于风格迁移，无需任何进一步的训练或修改。

为验证所提方法的鲁棒性与通用性，对于数据集使用 25 张输入源图像与 55 种不同风格。源内容图像来自 ImageNet[129] 及网络收集，涵盖风景、动物、交通工具、食物、建筑等多个领域。所用风格包括 32 种在实际中难以找到合适的参考图像地想象类风格和 23 种常见类风格。其中想象类风格，如“混沌画作（Painting of Chaos）”，往往没有对应的风格图像，难以用图像示例来表示。通过融合两种常见类风格，共设计了 32 种想象类风格，其中包括深海哥特神秘风格（Deep Sea Gothic Mystery）和丛林蒸汽朋克（Jungle Steampunk）。而对于 23 种常见类风格，由于较为方便的找到对应的风格参考图像，因此加入了与图像引导方法的对比。数据集中源图像可见图 4.2，想象类风格和常见类风格的示例可见表 4.1 和图 4.3。在与其他方法的对比上，对比了 11 种最先进（State-Of-The-Art，SOTA）的风格迁移方法，包括 8 种文本引导的风格迁移或图像编辑方法：SDXL[72]、CLIPstyler[125]、DiffuseIT[71]、ZeCon[73]、PTI[77]，InfEdit[78]、Freestyle[74] 以及 PnP Inversion[79]，同时也包含 3 种图像引导的风格迁移方法：Stytr2[60]，CAST[58]，以及 InST[63]。在对比实验中使用这些方法的公开实现，方法中可调整的配置均使用各方法推荐的配置。

图 4.4 利用本章提出的方法进行文本引导的图像风格迁移结果示例

图 4.4 展示了本章方法在想象类风格和常见类风格下均获得了不错的效果。

4.2.2 想象类风格的对比实验

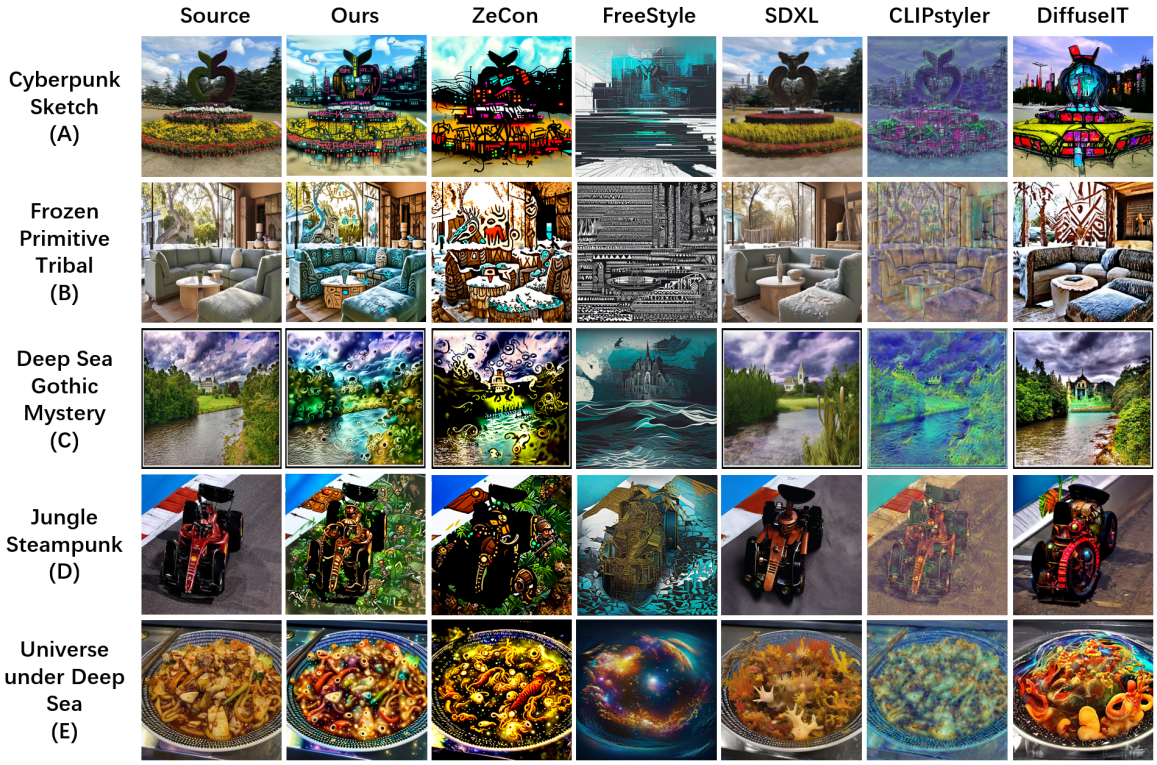


图 4.5 本章提出的方法与其他文本引导的风格迁移方法在想象类风格下的定性比较

本章所考虑的风格迁移方法在场景设置中基于文本引导，使用户能够通过文本描述来表达新颖的、想象类的风格，并将其迁移到源内容图像上。针对这些想象类风格，选用 5 个文本引导的风格迁移方法进行比较，如图 4.5 所示。

主要面向图像编辑任务设计的 DiffuseIT[71]，在应用至风格迁移任务时，通常会保留输入源图像的风格，而仅对其内容进行部分修改。例如，行 D 中源图像中的赛车被改造成带有树木元素的蒸汽朋克风格，以贴合丛林（Jungle）主题，但整体的风格化转变并不明显，导致风格化不足。CLIPstyler[125] 在风格化过程中会同时修改整张图像的色调和纹理，有点类似于加滤镜的效果。虽然这种方法能保证整体风格的一致性，但对内容的风格化幅度有限。

在对 SDXL 的图生图（img-to-img）设置中[72]，将强度（Strength）调整为 0.6，并将无分类器引导尺度（Classifier-free Guidance Scale）设为 7.5。以上两个参数分别用于控制图像生成过程中对原始图像进行修改的程度以及调节生成过程中对文本提示词依赖的强度。SDXL 在如行 A、C、D 和 F 的情况下会显著改变图像内容，但在

行 B 和行 E 中，生成的结果与源图几乎无差别。这说明 SDXL 对参数设置较为敏感，且常常需要反复尝试才能为不同的内容和风格提示词找到合适的参数。与 DiffuseIT 相似，SDXL 也仅对图像的部分内容进行局部修改，较难对图像进行全局风格化。

FreeStyle[74] 有时在风格化过程种会几乎完全丢失内容信息，如行 A 和 C 所示，有时则仅保留内容的大致轮廓，却缺乏任何与源内容相关的其他语义细节，如行 B、D、E 和 F 所示。ZeCon[73] 则既对图像中内容进行风格化，又在一定程度上保留了源图像的色彩基调。但它也会偶尔遗漏一些源内容细节，例如在行 B 中窗外房子的细节，以及在行 D 中赛车前悬挂系统的部分元素。

比之下，本章提出的方法展现了更优的平衡，既能对源图内容进行有效的风格化转换，又能较好地保留其轮廓及关键语义信息。这主要得益于本方法在设计上的两个关键方面。首先，在风格控制上，滑动窗口裁剪（SWC）与测地曲面特征增强（FAGS）的结合促进了全局信息融合，并通过在预形状空间中定义的风格损失实现了更准确的风格引导。其次，在内容保持上，预形状自相关一致性（PSC）模块有效约束了内容结构的稳定性。例如，在行 F 中，本章方法将米粉转变为带有星星点缀的深海生物，在行 B 中，则把赛车照片处理成具有蒸汽朋克质感的画面。此外，本章方法保证了整张图像的风格化一致性。比如，行 C 中可以看到本章方法生成的图像能够对天空、森林和水域进行无缝融合，将其转变为融合哥特风格元素的深海风格，并保留了建筑物的细节。

表 4.2　想象类风格的定量比较。Ours 表示本章方法、FS 表示 FreeStyle、CS 表示 ClipStyler、DIT 表示 DiffuseIT。最佳结果用加粗表示，次优结果用下划线标注。

指标	Ours	ZeCon	FS	SDXL	CS	DIT
PSNR ↑	<u>28.27</u>	27.94	27.95	**28.82**	27.89	28.21
SSIM ↑	**0.498**	0.277	0.138	<u>0.440</u>	0.297	0.202
LPIPS ↓	**0.314**	0.539	0.695	<u>0.341</u>	0.458	0.440
CLIP-I ↑	**0.334**	<u>0.317</u>	0.268	0.224	0.273	0.252
CLIP-P ↑	**0.242**	<u>0.239</u>	0.218	0.211	0.226	0.216

为更全面地评估本章方法的性能，使用了多种定量指标，如表 4.2 所示。在对方法源图内容保持能力的评估上，计算了峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）、结构相似性（Structural Similarity Index Measure，SSIM）[130] 与学习感知图像块相似度（Learned Perceptual Image Patch Similarity，LPIPS）[106]，以衡量生成图像与源图像

在内容上的相似度。在对方法的风格化能力评估上，计算了 CLIP 分数来评估生成图像与风格参考文本经过 CLIP 提取得到的特征之间的风格相似度。CLIP-I 针对整张图像进行计算，反映了整体的风格迁移程度，CLIP-P 则在多张 64×64 的图像块上取平均，从而衡量图像在局部上的风格一致性。结果显示，本章提出的方法在 SSIM、LPIPS、CLIP-I、CLIP-P 指标上均取得了最高分，说明在源图内容保持与风格化能力与一致性上具有领先表现。SDXL 善于生成高质量图像，因此它的 PSNR 略高于本章方法，但本章方法生成图像的 PSNR 仍能排在第二位。在 CLIP 分数的对比上，本章方法仅微幅领先了 ZeCon，可见这两种方法在风格化能力方面都具备较高的水准。然而，ZeCon 在内容保持上有所不足，其 PSNR 与 SSIM 得分均不及本章方法。综合而言，实验结果表明，本章方法能较好地兼顾生成图像的风格化程度、整体图像风格一致性以及对源图中内容的保持。

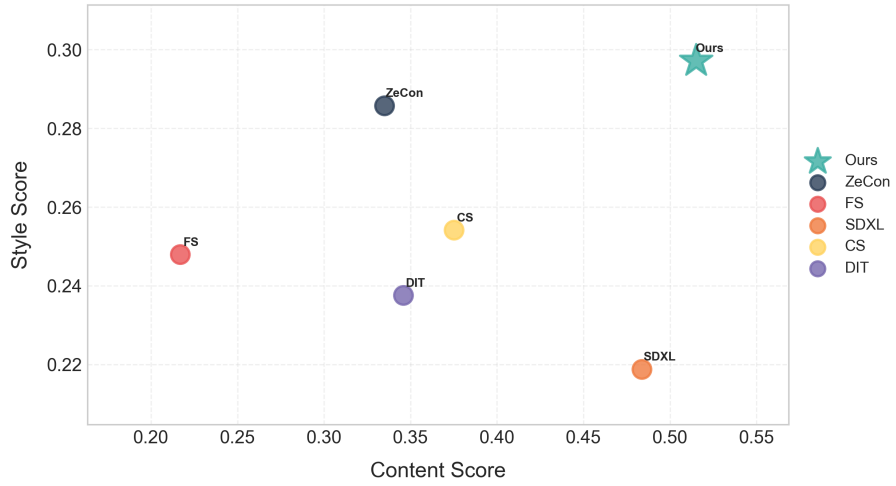


图 4.6 想象类风格下内容保真度（横轴）与风格一致性（纵轴）的二维比较。两个轴上的值越高，表示源内容的保留效果越好，风格化越均匀。

为了在基于表格的比较之外进一步说明本章方法的整体优越性，此处展示了一个二维散点图，该图将多个指标整合为“内容分数”（横轴）和“风格分数”（纵轴）。本章方法将内容相关指标 PSNR、SSIM 和 LPIPS 合并为单一的内容分数，并将风格相关指标 CLIP-I 和 CLIP-P 合并为风格分数。具体而言，首先将每个指标归一化到 [0,1] 区间，其中 1 表示最佳值。由于 PSNR 通常介于 20 dB 到 60 dB 之间，将其线性映射到 [0,1]。SSIM、CLIP-I 和 CLIP-P 本身就在 [0,1] 区间内，对于 LPIPS，则通过取 1 减去其值进行转换。完成这些调整后，分别以 0.2、0.4 和 0.4 的权重对 PSNR、

SSIM 和 LPIPS 进行加权，构成内容分数轴。PSNR 虽然提供了对信号失真的直接度量，但其与感知的对齐程度不如 SSIM 和 LPIPS，后两者更能可靠地捕捉微妙的结构和感知变化。对于风格分数轴，CLIP-I 和 CLIP-P 以 0.6 和 0.4 的权重聚合，更侧重于整体风格对齐而非局部图像块的一致性。较高的内容分数表示对源图像结构和细节的更精确保留。相对地，较高的风格分数表示在整个图像上更均匀和忠实的风格化。散点图中越靠近右上角的位置表明在保留原始内容和实现均匀风格迁移之间达到了更理想的平衡。

如图 4.6 所示，本章方法的分数显著位于右上区域，在两个维度上均超越了所有其他方法。虽然 SDXL 获得了值得称赞的高内容分数，但其风格分数相对落后，表明风格化应用得不够全面或一致。另一方面，诸如 FS 和 DIT 等方法则位于图的左下部分，暗示其在内容保真度或风格均匀性方面存在权衡。这些结果表明，本章的方法能较好地兼顾生成图像的风格化程度、整体图像风格一致性以及对源图中内容的保持。

4.2.3 常见类风格的对比实验

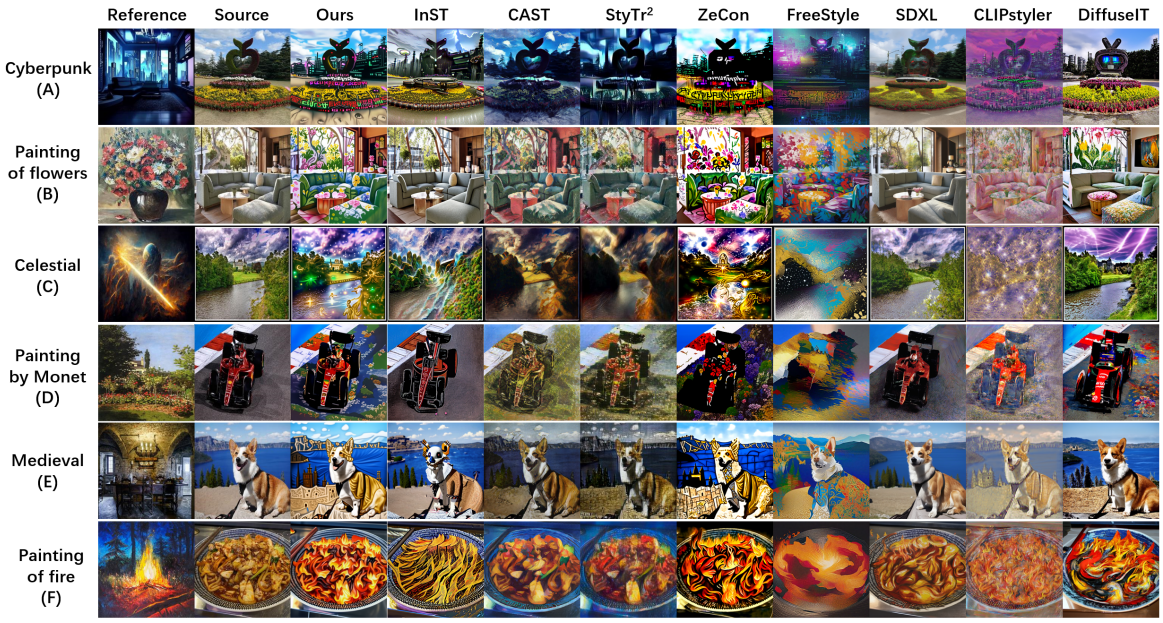


图 4.7 本章方法与其他风格迁移方法在若干常见类风格上的定性比较。

本小节选用了 23 种常见类风格，如赛博朋克（Cyberpunk）、中世纪（Medieval）等来进行实验。由于常见类风格可在网上找到较为贴合文本提示的参考风格图像，在

本小节不仅对比了 5 种文本引导方法，还额外加入了 3 种图像引导的方法，对于文本引导方法仍使用文本提示词进行引导，对于图像引导方法则使用文本对应的风格图像进行引导。图 4.7 展示了本章提出的方法与这 8 种风格迁移方法的定性结果。

对常见类风格的处理，文本引导方法与之前处理想象类风格时面临的挑战相似。比如，DiffuseIT 和 SDXL 有时会过度修改内容，如行 A 所示，有时却几乎没有变化，例如行 E 中的柯基变化极其细微、行 B 则仅把桌面物件变成花瓶，或把窗外风景变成花。至于 ClipStyler，虽然会更改整张图像的色调和纹理，但整体画面常呈现灰暗而模糊的效果，与源图的色调差别过大。相比之下，FreeStyle 仅保留了内容的大致轮廓，并直接将相应风格填充到这一轮廓中。ZeCon 则在保留图像语义信息的同时进行风格化，如行 B 中可辨识的沙发与桌子，但有时会丢失内容的部分轮廓例如行 A 中植物雕塑的外形以及 C 中房屋的整体结构都产生了偏移或丢失。

在图像引导的对比方法中，会选取与文本提示尽量匹配的参考图像，以进行公平的风格迁移。StyTr2[60] 和 CAST[58] 都是通过从参考图像中提取解耦的纹理和色彩信息，并融合至源图中来执行风格化。但只依赖单张图像中包含的纹理和色彩等信息难以完整呈现文本中所包含的丰富风格特征。以行 A 的赛博朋克风格为例，StyTr2 和 CAST 仅把源图改成蓝黑色调，而没能充分捕捉此风格中的建筑元素或霓虹灯色彩等要素。但此类方法由于有图像作为参考，它们在和油画、色调、纹理等密切相关的风格上能取得较好效果，如行如图 B、D 和 F 所示，但对其他更复杂的风格时有所欠缺。InST[63] 则先利用文本反演（Textual Inversion）将参考图像转换为文本嵌入，再结合 Stable Diffusion 来进行风格迁移。由于它额外融合了来自图像的文本信息，能在行 A 的 Cyberpunk 迁移中把树林变成该风格的建筑。但在行 D 和行 E 中，InST 生成图像的风格统一度仍不够，往往只对图中主体部分，如赛车或柯基，的风格进行了迁移，而忽略了其他部分。

与以上方法相比，本章方法更能在保留原图内容的基础上，对每个部分都进行有效且一致的风格化。例如，在行 A 中，本章方法将森林变换成了赛博风格的建筑，又如行 B，源图中沙发的轮廓得以保留，同时增添鲜明的花朵纹样。更重要的是，本章方法能够让整张图的风格保持一致化。以行 E 为例，不仅源图中的主体柯基变为了中世纪风格，周围的森林、地面及远处山体都相应地转变为相同风格，从而在全局范围内呈现出更融合的风格效果。

在这些常见类风格的对比实验里，选取全部对比的风格迁移方法进行了定量分

表 4.3　常见类风格下的定量比较。Ours 表示本章方法、FS 表示 FreeStyle、CS 表示 ClipStyler、DIT 表示 DiffuseIT。表格中最佳结果以加粗标识，次优结果则下划线标识。

指标	Ours	InST	CAST	Stytr^2	ZeCon	FS	SDXL	CS	DIT
PSNR ↑	28.29	**28.52**	27.98	27.98	28.07	27.96	<u>28.45</u>	27.89	28.18
SSIM ↑	**0.510**	0.389	0.402	<u>0.415</u>	0.356	0.136	0.369	0.295	0.182
LPIPS ↓	**0.308**	0.409	0.474	0.526	0.489	0.692	<u>0.322</u>	0.463	0.455
CLIP-I ↑	**0.293**	0.229	0.242	0.237	<u>0.292</u>	0.260	0.252	0.266	0.252
CLIP-P ↑	**0.236**	0.212	0.229	0.223	<u>0.234</u>	0.219	0.222	0.233	0.219

析，其结果如表 4.3 所示。和想象类风格的结果相似，本章方法在 SSIM、LPIPS 与 CLIP 指标上都获得了最高分，说明本章方法生成的图像既能保留图像的内容，也能保证足够程度的风格化以及风格一致性。在 PSNR 指标上，本章方法的结果略低于分列第一与第二的 InST 和 SDXL，这两种方法都基于 Stable Diffusion 作为骨干网络，因而成像质量较高。然而，在风格指标上，这些方法未能超越本章方法。本章方法和 ZeCon 仍然是在风格化效果上最佳的两个方法，见表中本章方法的 CLIP 分数领先，第二名即为 ZeCon，然而 ZeCon 在内容保留方面相比略显不足。

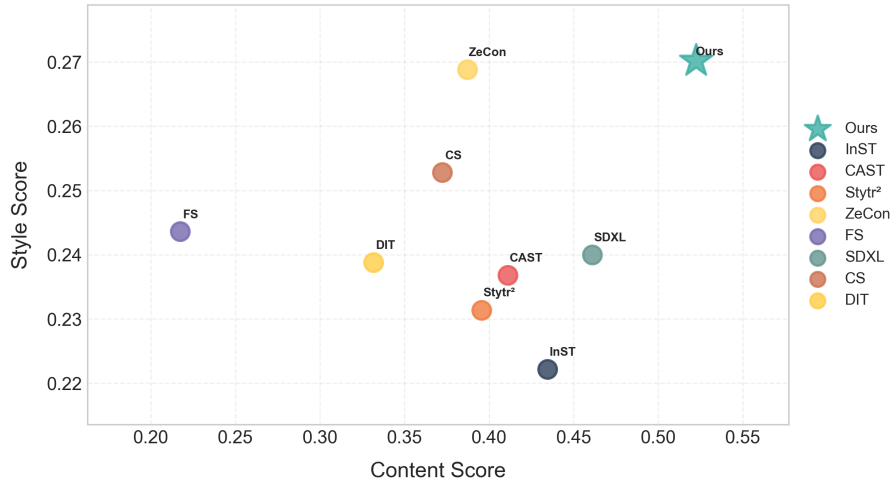


图 4.8　常见类风格下内容保真度（横轴）与风格一致性（纵轴）的二维比较。两个轴上的值越高，表示源内容的保留效果越好，风格化越均匀。

除了表格结果外，本节还将这些常见类风格的比较结果可视化在一个二维散点图中，如图 4.8 所示。每种方法根据其内容分数和风格分数进行定位，便于观察其在内容保留和风格化效果之间的平衡。虽然 InST 和 SDXL 显示出较高的内容保真度，但它们的风格一致性并未超过本章方法。同时，ZeCon 的位置反映了其良好的风格

化能力，但内容相似度相对较低。本章方法占据了右上区域，展示了本章方法在多种风格下实现高保真且风格一致的图像转换的优越能力。

4.2.4 与图像编辑方法的对比实验

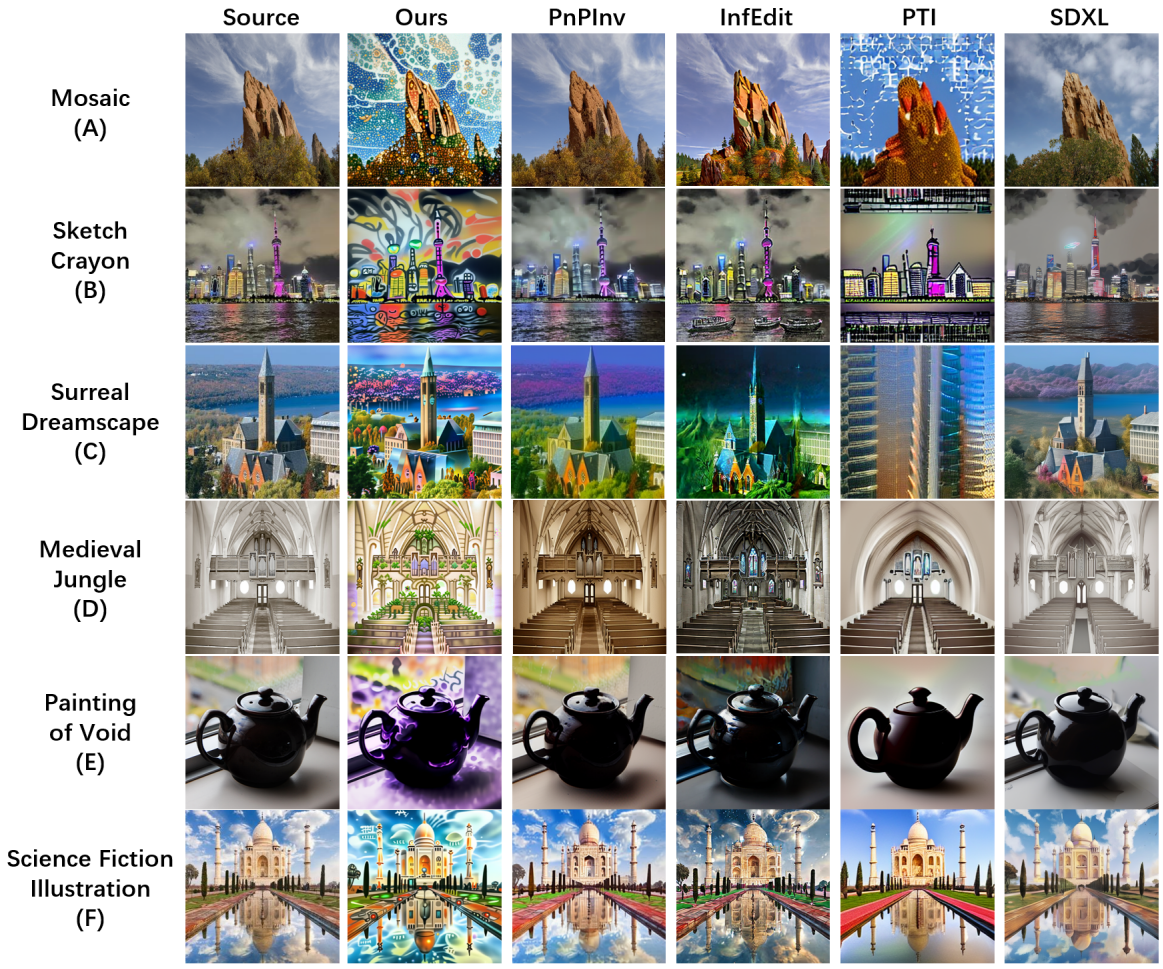


图 4.9 本章方法与其他图像编辑方法在想象类和常见类风格下的定性比较

本节评估了本章提出的方法与 4 种基于扩散模型的图像编辑方法，包括 PnPInv[79]、InfEdit[78]、PTI[77] 和 SDXL，在想象类和常见类风格下的表现。本节仔细调整了每种图像编辑方法的参数以实现更强的风格化效果。例如，提高了 guidance_scale（引导尺度）、cross_replace_steps（交叉注意力替换步数）和 self_replace_steps（自注意力替换步数）来进一步强化风格[131]。此外，还使用了比本章方法自身所用提示词更详细的提示词（prompt）。对于一块以岩石为主体的源图像，首先将源提示词标记为“一块岩石（a rock formation）”，然后在运行每种方法时将目标提示词转换为“[马

赛克] 风格的一块岩石（a rock formation in [Mosaic] style）”。图 4.9 展示了定性比较结果，说明了每种技术在诸如马赛克（Mosaic）和科幻插画（Science Fiction Illustration）等风格化提示词下如何处理相同的源图像。

在图 4.9 中，本章方法始终能在保持原始图像内容的同时实现清晰的风格化，例如，行 A 中细节丰富的马赛克效果和行 D 中清晰的中世纪丛林元素。PnPInv 和 SDXL 保持了结构的清晰度，但只提供了微小的风格变化，如行 D 和行 E 所示。InfEdit 在行 C 中表现良好，创造了一个符合超现实梦境（Surreal Dreamscape）风格的梦幻场景。然而，它在行 A 中的风格强度不足，并且在行 B 中引入了源图像中没有的额外内容（许多船只）。PTI 在内容清晰度和风格强度方面都存在问题，在行 A 和行 B 中显示出模糊的形状和不完整的风格化。本章方法在应用强风格的同时保持源图像关键结构的完整性方面表现最为一致。

表 4.4　本章方法与其他图像编辑方法在想象类和常见类风格下的定量比较。Ours 表示本章方法。最佳结果以加粗表示，次优结果用下划线标注。

指标	Ours	PnPInv	InfEdit	PTI	SDXL
PSNR ↑	<u>28.28</u>	**28.60**	28.09	28.26	28.23
SSIM ↑	<u>0.503</u>	**0.585**	0.422	0.358	0.452
LPIPS ↓	<u>0.312</u>	**0.223**	0.428	0.487	0.335
CLIP-I ↑	**0.311**	0.224	<u>0.236</u>	0.227	0.232
CLIP-P ↑	**0.236**	0.210	0.214	0.208	<u>0.215</u>

表 4.4 展示了本章方法与其他图像编辑方法在想象类和常见类风格下的各项数值得分。本章方法在 CLIP-I 和 CLIP-P 上得分最高，表明其能紧密贴合风格提示词，并在整个图像上保持风格的一致性。PnPInv 在内容保真度指标上领先，这意味着它很好地保持了图像质量，但在 CLIP 分数上较低，表明其风格应用程度较低。InfEdit 和 PTI 在特定指标上表现良好，但无法匹敌本章方法在风格应用上的一致性。SDXL 在大多数指标上居中，有时能提供良好的细节，但其较低的 CLIP-P 结果显示它未能达到相同的整体风格水平。

为了更直观地比较，本节还将每种方法的分数可视化在一个二维散点图中，如图 4.10 所示。虽然 PnPInv 取得了较高的内容保真度分数，但其风格化能力较弱。正如在图 4.9 的定性结果所示，PnPInv 通常能很好地保留原始内容，但仅轻微地应用

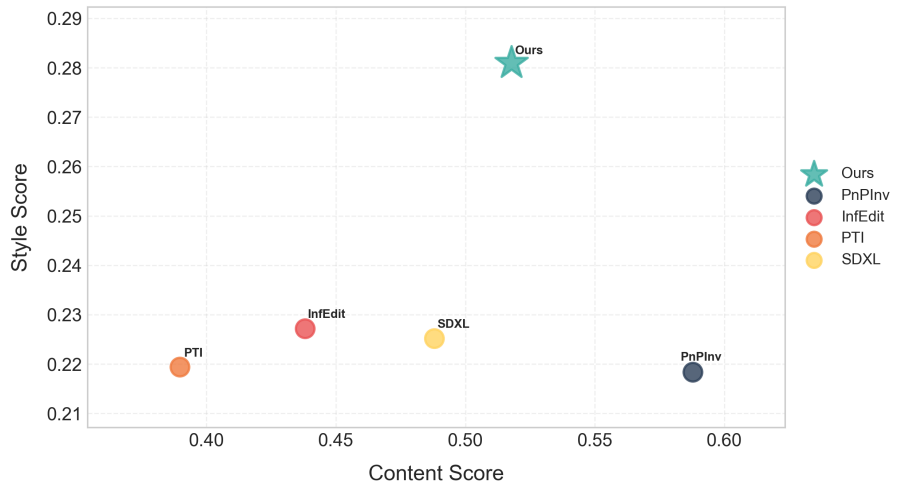


图 4.10 在想象类和常见类风格下，内容保真度（横轴）与风格一致性（纵轴）的二维比较。两个轴上的值越高，表示源内容的保留效果越好，风格化越均匀。

了目标风格。InfEdit 的得分位于中间区域。PTI 和 SDXL 的分数在纵轴上偏低，表明风格应用不够彻底。相比之下，本章方法的分数出现在右上区域附近，显示出在内容保留和风格完整性两方面都具有强大的性能。本章方法在内容保真度和风格一致性之间实现了更优的平衡（Trade-off），确认了其在所测试方法中的整体优势。

4.2.5 消融实验

为验证本章中提出的基于 FAGS 模块的特征融合以及预形状自相关一致性（PSC）模块的有效性，本小节选取了一个苹果形状的植物雕塑作为源内容图像，结合 5 种不同风格，进行了定性和定量的消融实验。在风格控制损失中，分别引入滑动窗口裁剪（SWC）以及 FAGS 特征融合，以促进图像块间的信息交互，确保风格信息的均匀一致迁移。此外，在内容控制损失中，将图像特征投影至预形状空间并计算 PSC 模块，以更好地平衡风格化与内容保真度。

为了提高风格迁移的一致程度，本研究将 ZeCon 中使用的随机裁剪替换为 SWC 策略。基线方法 ZeCon 采用随机裁剪，在多种风格下，如蜡笔素描、马赛克、水彩和浮世绘，常常导致雕塑形状信息的丢失。此外，ZeCon 在整体风格转换上表现出不一致性。例如，在立体主义（Cubism）风格化结果中，雕塑被风格化了，而背景中的树林却保持不变。将裁剪图像块的策略切换到 SWC 能更有效地保留内容，尽管最初会导致轻微的风格化不足，如图 4.11 和表 4.5 的第 2 行所示。通过增加风格控制损失的权重，图 4.11 的第 3 行展示了对源图像内容更好的保留效果以及相比基线更

图 4.11　提出的风格控制改进方法的定性消融研究。其中 * 表示将风格损失权重 λ_{sty} 从默认值 20000 提升至 25000。

表 4.5　提出的风格控制组件的定量消融研究。最佳结果以加粗表示，次优结果用下划线标注。S 和 F 分别表示 SWC 和 FAGS。其中 * 表示将风格损失权重 λ_{sty} 从默认值 20000 提升至 25000。

编号	模块		指标				
	S	F	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-I ↑	CLIP-P ↑
1	✗	✗	28.112	0.469	0.407	<u>0.253</u>	0.221
2	✓	✗	**28.274**	**0.537**	**0.311**	0.203	0.211
3*	✓	✗	28.201	0.473	0.402	0.244	0.221
4	✗	✓	28.145	0.476	0.387	**0.259**	<u>0.230</u>
5	✓	✓	<u>28.223</u>	<u>0.495</u>	<u>0.386</u>	<u>0.253</u>	**0.223**

强的风格应用。这也可以从表 4.5 中增加的 PSNR 和 SSIM 以及相同的 CLIP-P 分数反映出来。引入 SWC 后，无论是背景树林还是雕塑主体都能在立体主义风格下保持风格一致性。

如图 4.11 第 3 行所观察到的，应用更高的风格损失权重有时会导致图像不同部分之间的不一致，例如水彩风格化中树林左右两侧的不一致。此外，增加权重往往会牺牲内容的保留效果，这在马赛克和水彩风格化图像中雕塑颜色信息的丢失上得到了证明。

这表明，即使引入了 SWC，原始的风格控制损失仍需使得图像块之间有足够的信息交互，导致在较高权重下出现不均匀的风格化和部分内容信息的丢失。为了解决这些问题，本章方法加入了 FAGS 模块以增强图像块之间的交互。如图 4.11 第 4 行所示，加入 FAGS 使得背景风格化更均匀，并且更好地保留了雕塑的颜色和结构。在定量分析方面，表 4.5 的第 4 行显示，相比于使用更高 λ_{sty} 的 SWC 版本，其 CLIP 分数有所提高，LPIPS 分数更低，表明风格一致性和内容保留效果更好。结合 SWC 和 FAGS 进一步提升了所有风格下的视觉质量，如第 5 行所示。

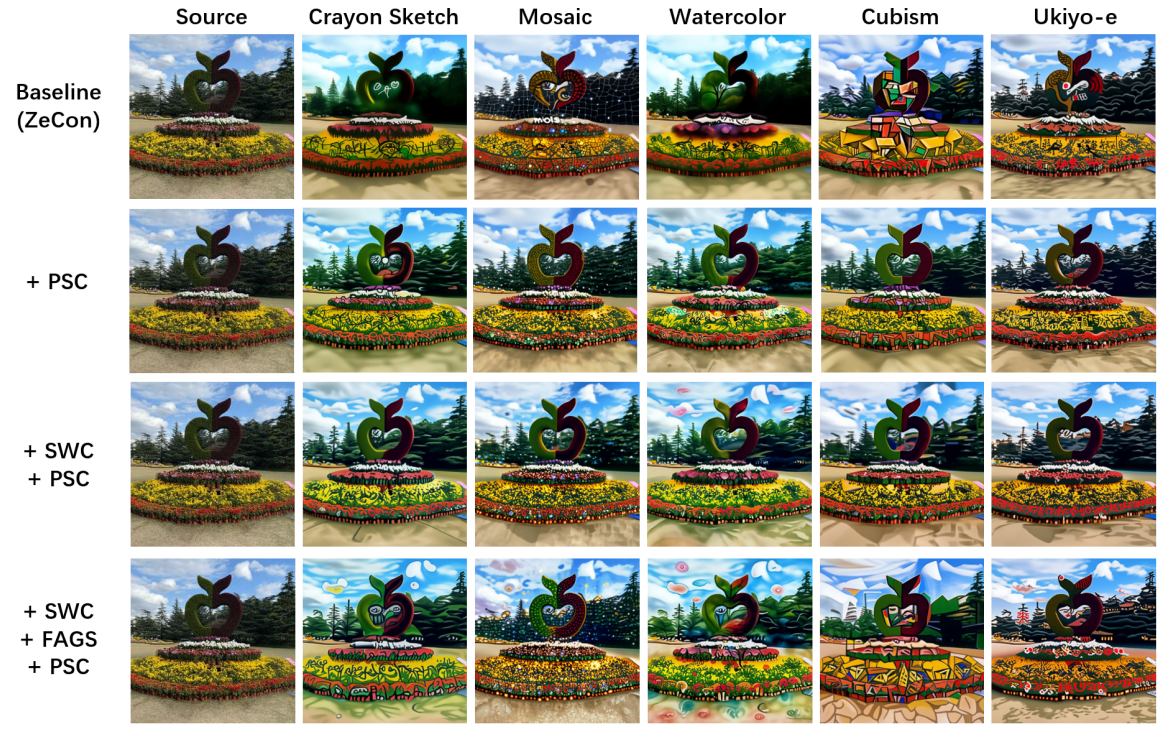


图 4.12　本章方法所提出的预形状自相关一致性模块的定性消融研究

SWC 和 FAGS 策略显著解决了风格不一致的问题，并增强了内容中颜色信息的

保留。然而，这些策略可能会影响内容轮廓形状的保留。例如，在图 4.11 第 5 行显示的立体主义风格化中，雕塑底座呈现不完整的现象。

为了更好地保留源图像内容中的形状，本章方法引入了预形状自相关一致性（Pre-Shape Self-correlation Consistency，PSC）模块。加入 PSC 后，生成的结果在所有风格下都能更好地保持雕塑的结构。例如，在图 4.12 第 2 行的蜡笔素描和马赛克风格中，雕塑的轮廓和内部颜色区域得以保留，而相比之下，第 1 行的基线中，风格化常常会使内容形状出现扭曲。

表 4.6 本章方法中提出的预形状自相关一致性模块的定量消融研究。S、F 和 P 分别表示 SWC、FAGS 和 PSC。

编号	模块			指标				
	S	F	P	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-I ↑	CLIP-P ↑
1	✗	✗	✗	28.112	0.469	0.407	0.253	0.221
2	✗	✗	✓	28.298(+0.186)	0.556(+0.087)	0.313(-0.094)	0.217	0.217
3	✓	✗	✓	28.294(+0.182)	0.548(+0.079)	0.326(-0.081)	0.214	0.216
4	✓	✓	✓	28.238(+0.126)	0.498(+0.029)	0.339(-0.068)	0.252	0.223

在没有任何风格增强模块的情况下，仅 PSC 本身就显著改善了内容保留效果，如表 4.6 所示，PSNR (+0.186)、SSIM (+0.087) 和 LPIPS (-0.094) 均有提升。当与 SWC 和 FAGS 结合使用时，如第 4 行所示，结果进一步改善。在水彩和浮世绘等风格中，雕塑形状保持良好，背景也被更均匀地风格化。与基线相比，最终版本在所有内容相关指标上均表现更优，同时也取得了具有竞争力的 CLIP-I 分数和最高的 CLIP-P 分数。

为了评估内容控制损失 L_{cont} 中每个项的效果，本小节选择性地移除各个损失分量，同时保留其他分量，进行消融分析。如表 4.7 所示，L_{cont} 包括四个部分：L_{psc}、L_{ZeCon}、L_{VGG} 和 L_{MSE}。在表中的各个情况下，L_{psc} 都被保留作为基础模块，研究重点是比较其余各项的贡献。

观察发现，移除 L_{ZeCon} 会导致内容相似度显著下降：SSIM 降低了 0.152，LPIPS 增加了 0.153。移除 L_{VGG} 时也观察到类似趋势，尽管影响稍小。这些结果表明，L_{ZeCon} 和 L_{VGG} 在内容保真度方面都发挥着重要作用，其中 L_{ZeCon} 对感知相似度的贡献更大。当同时移除 L_{ZeCon} 和 L_{VGG} 时，模型在所有指标上表现均不佳，显示它们是互补

表 4.7 内容控制损失中各组成项的定量消融研究

内容损失	指标				
	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-I ↑	CLIP-P ↑
完整 L_{cont}	28.239	0.498	0.339	0.252	0.223
移除 L_{ZeCon}	27.906(-0.333)	0.346(-0.152)	0.492(+0.153)	0.256	0.230
移除 L_{VGG}	27.790(-0.449)	0.342(-0.156)	0.487(+0.148)	0.263	0.232
移除 L_{MSE}	28.296(+0.057)	0.550(+0.052)	0.328(-0.011)	0.215	0.213
移除 L_{MSE}, L_{VGG}	27.937(-0.302)	0.393(-0.105)	0.402(+0.063)	0.231	0.226
移除 L_{ZeCon}, L_{VGG}	27.775(-0.464)	0.252(-0.246)	0.539(+0.200)	0.248	0.229
移除 L_{ZeCon}, L_{MSE}	27.768(-0.471)	0.238(-0.260)	0.541(+0.202)	0.248	0.230

的。移除 L_{MSE} 会轻微提高 PSNR 和 SSIM，这表明虽然它改善了像素级对齐，但它不出现在内容控制损失中使得网络能够更专注于结构级的一致性。然而，内容分数的提高是以较低的 CLIP 分数为代价的，这表明像素对齐对于保持风格相关性较为重要。

包含所有四个组成项的完整内容损失，在内容保留和风格兼容性之间实现了最佳平衡。这些结果证实了最初在 ZeCon 中提出的所有其余项对内容保持做出了互补的贡献。本章方法新引入的 L_{psc} 是一个额外的模块，进一步加强了形状级的内容一致性。

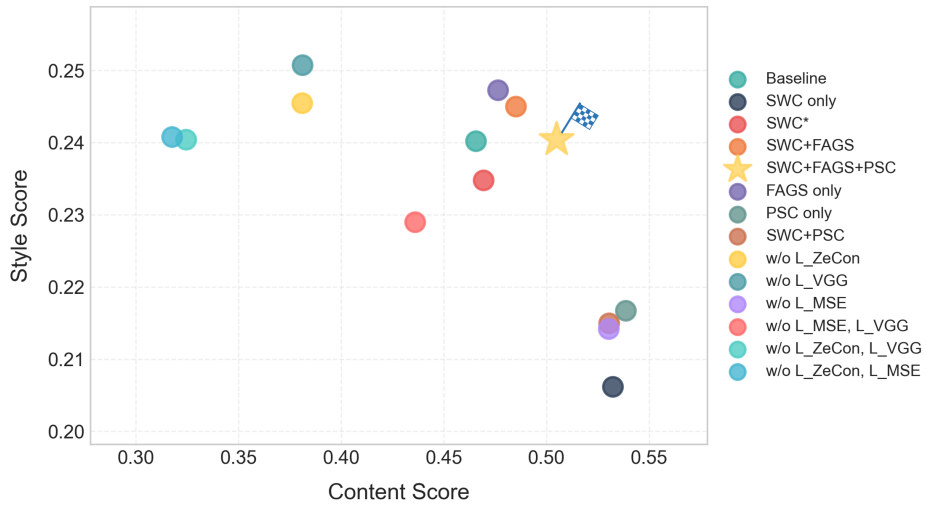


图 4.13 内容保真度（横轴）与风格一致性（纵轴）的二维消融比较。两个轴上的值越高，表示源内容的保留效果越好，风格化越均匀。

为总结上述各个模块的消融研究，本节将本章方法的所有变体可视化在一个二

维散点图中，如图 4.13 所示。每个点代表一个方法变体，横轴表示内容保真度，纵轴表示风格一致性。

虽然像“仅 FAGS”或“SWC+FAGS”这样的模块取得了较高的风格分数，但它们在内容保真度方面有所欠缺。相反，“仅 PSC”改善了内容保留，但缺乏足够的风格化。如图 4.11 和图 4.12 中的定性结果所示，PSC 模块在维持被其他组件经常扭曲或丢失的内容结构方面起着关键作用。在图 4.13 的下半部分，可以看到移除内容损失的任何一个分量都会导致内容保真度的明显下降，证实了它们各自都发挥着重要作用。结合了 SWC、FAGS 和 PSC 的本章方法，其分数最接近右上角，显示了在内容保留和风格化之间的最佳权衡。图 4.13 验证了集成所有提出组件的重要性，其中 SWC 用于更好的区域覆盖，FAGS 用于图像块级的风格一致性，PSC 用于形状级内容引导，以及完整的内容损失用于稳定的内容保持。

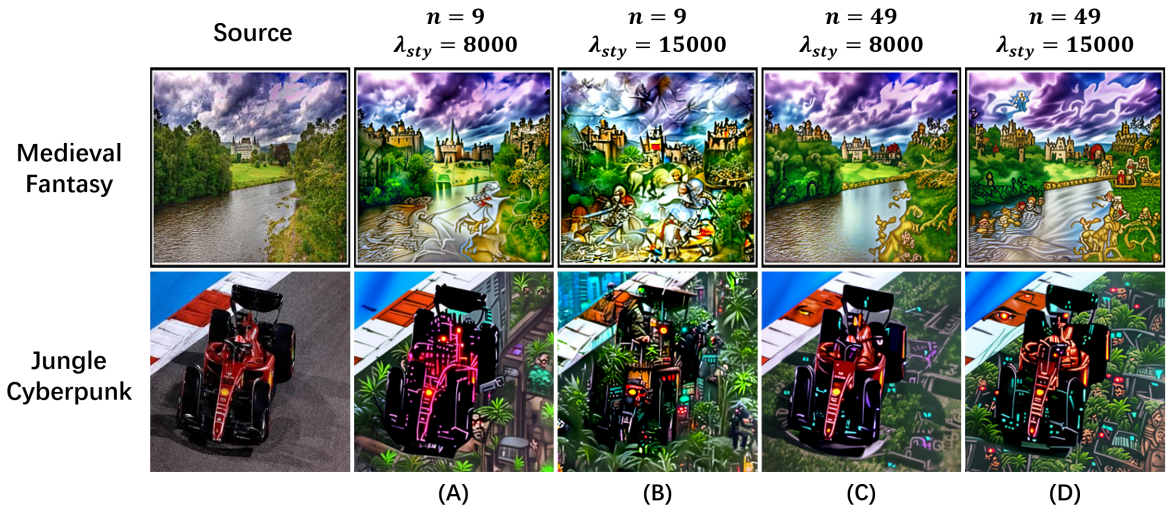


图 4.14 超参数图像块数量 n 的消融实验。第 1 行对应中世纪幻想（Medieval Fantasy）风格，第 2 行对应丛林赛博朋克（Jungle Cyberpunk）风格。

在本章提出的方法中，共需设置 7 个超参数，包括 SWC 所裁剪出的图像块数 n 以及风格控制损失 L_{pc}、L_{pd} 和内容控制损失 L_{psc}、L_{ZeCon}、L_{VGG}、L_{MSE} 中各项的权重。

图 4.14 展示了不同图像块数 n 对风格迁移结果的影响。当小块数仅为 9 时，风格控制损失的权重会对图像质量产生明显影响，如列 B 中的中世纪风格看起来杂乱无章，并导致苹果形雕塑的轮廓丢失。在较低权重设置下，也会出现内容缺失问题，如列 A 中赛车座舱轮廓不完整。

反之，若 $n = 49$，每个裁剪得到的图像块的大小为 32×32，更小尺寸的图像块有助于更细致地保留内容。此时，可通过调整风格控制损失的整体权重来控制风格化强度，但对内容保留几乎不造成负面影响，因而能在保持原图完整性的同时，更自由地实现预期风格效果。

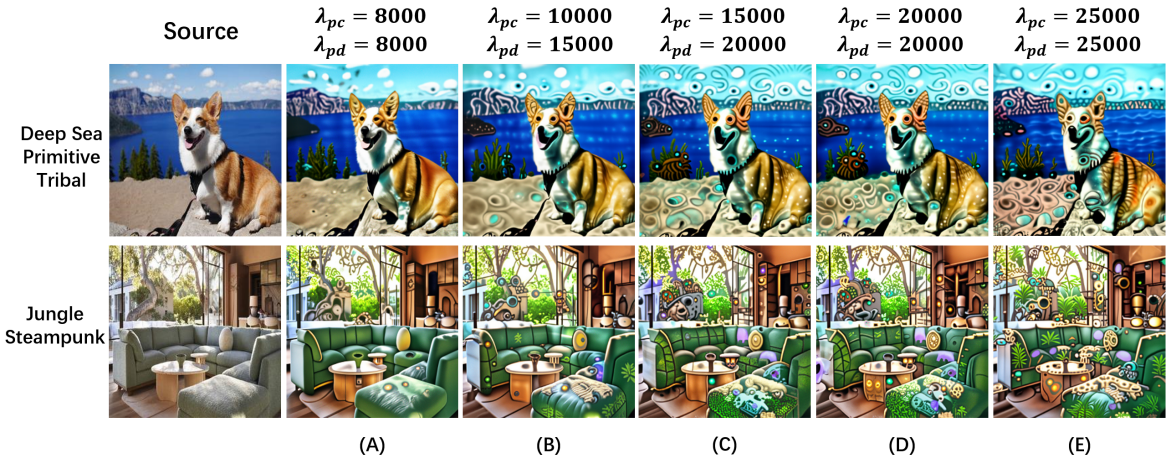


图 4.15　超参数 λ_{pc} 与 λ_{pd} 的消融实验。第 1 行对应深海原始（Deep Sea Primitive Tribal）风格，第 2 行对应丛林蒸汽朋克（Jungle Steampunk）风格。

如图 4.15 所示，若将 L_{pc} 和 L_{pd} 的权重 λ_{pc} 与 λ_{pd} 设得过低，如小于 15000，容易出现风格化不足的情况。经过大量实验发现，把以上权重都设在 15000 至 25000 范围内能在多种源图像与风格提示词下取得理想效果。故本章默认将 λ_{pc} 和 λ_{pd} 都设为 20000。

表 4.8　本章方法的超参数默认设置

超参数	n	λ_{pc}	λ_{pd}	λ_{psc}	λ_z	λ_v	λ_m
值	49	20000	20000	1000	1000	1000	100

在原有内容控制损失部分，与 ZeCon[73] 的超参数保持一致，将 L_{ZeCon}、L_{VGG}、L_{MSE} 的权重分别设为 1000、1000 和 100，并统一将 L_{psc} 的权重 λ_{ps} 设为 1000。表 4.8 列出了本章方法所有超参数的默认设置。

4.3　本章小结

本章提出了名为 FAGStyle 的零样本文本引导扩散式图像风格迁移方法。通过将滑动窗口裁剪技术与测地曲面上的特征增强相结合引入风格控制损失中，方法在风

格一致性上得到显著提升，促进不同图像区域间信息的交互，进一步验证了形状空间理论与测地曲面构建方法在多种图像生成模型架构以及多种图像生成任务的适配性。同时，为了保持内容一致性，引入了预形状自相关一致性，从而在视觉风格迁移的过程中有效保留原始内容结构。实验结果表明，FAGStyle 在多种风格场景中均具有较好的表现，相较当前主流方法能取得更高的风格保真度与更优的内容保留效果。

第五章　总结与展望

5.1　结论

深度学习驱动的图像生成技术，特别是生成对抗网络和扩散模型，在合成高逼真度图像方面取得了巨大成功，但普遍面临对大规模训练数据的依赖。在很多实际应用中，比如材料科学研究，数据稀缺性严重制约了这些模型的性能，导致生成图像质量下降、多样性不足及模式坍塌等问题。现有的小样本生成策略，如迁移学习和传统数据增强，在没有语义相关的大规模预训练模型或面对极端小样本的场景下，效果仍有局限，难以充分挖掘有限数据中的复杂结构信息。与此同时，在利用文本信息引导生成模型进行零样本风格迁移时，如何在准确注入新颖风格的同时，有效保持原始内容结构并避免语义失真，也对现有技术提出了严峻的挑战。

针对上述挑战，本论文的核心思想是创新性地引入了形状空间理论，利用其在去除无关变换后捕捉对象内在形态的能力，提出一种通用的、基于预形状空间的非线性特征增强策略。该策略通过分析和利用少量样本构建测地曲面，来进行有效的特征增强与插值，旨在利用数据的内在结构特性，为小样本生成中的信息挖掘和可控风格迁移中的信息融合与精确调控等信息受限场景，提供统一的技术支撑，进而提升生成模型的综合性能。

本论文的主要研究工作和贡献总结如下：

（1）提出了一种基于预形状空间测地曲面信息迁移的方法，应对极端小样本图像生成的挑战。该方法面向无源域、极端小样本场景，其核心在于将从判别器提取的有限样本特征投影到预形状空间，通过构建样本间的测地曲面进行非线性特征增强，这项工作通过提出的测地曲面特征增强模块实现，模拟更丰富的数据分布以构建伪源域，并将此信息有效迁移。同时，设计的插值监督与正则化模块进一步保证了生成过程的稳定性和生成图像的平滑过渡。实验证明，该方法在不依赖大规模预训练的前提下，显著提升了生成图像的质量和多样性，有效缓解了模式坍塌和过拟合问题。

（2）将预形状空间中测地曲面特征增强思想成功应用于文本引导的零样本图像风格迁移，实现了无需参考图或微调的高效文本控制与内容保持。本论文提出的方

法将基于预形状空间的测地曲面特征增强思想拓展到基于扩散模型的、更具挑战性的可控生成任务中。通过结合滑动窗口裁剪技术处理局部信息，并利用预形状空间中的测地曲面特征增强模块增强各区域特征间的交互与增广，有效提升了对文本描述风格的捕捉能力和全局风格一致性。同时，设计的预形状自相关一致性模块确保了在风格转换过程中对原始内容结构的稳定保持。实验表明，该方法无需风格参考图像或模型微调，即可实现灵活、高质量的零样本风格控制，在风格表达和内容保真度之间取得了良好平衡。

（3）通过系统的实验，验证了所提出的基于预形状空间测地特征增强策略的有效性及其在信息受限场景下的应用价值。实验设计涵盖了多种通用和特定领域的小样本及风格迁移数据集，从定性和定量角度，使用包括 FID、LPIPS、PSNR、SSIM 及 CLIP-score 在内的多种指标全面评估了所提方法。结果一致表明，基于预形状空间的测地曲面特征增强策略在提升生成图像的保真度、多样性与可控性方面具有显著效果。实验不仅验证了该策略有助于提升模型在小样本条件下的学习与生成能力，也证明了其在零样本条件下实现精确风格控制与内容保持的有效性，并展示了其在如材料图像数据增强等下游任务中的实际应用潜力。

5.2 工作展望

尽管本论文提出的基于形状空间理论的特征增强方法在小样本图像生成和零样本风格迁移任务上取得了一定的进展，但该研究方向仍有广阔的探索空间和诸多值得深入研究的方面:

（1）计算效率与可扩展性。基于测地曲面的特征增强计算相对复杂，尤其当样本数量或特征维度显著增加时，计算开销可能限制其在大规模场景下的应用。未来的工作可以致力于研究更高效的测地计算近似方法，例如探索基于图的测地近似、切空间线性近似的优化，或利用降维技术简化流形表示。

（2）可控性与生成质量的进一步提升。目前的方法主要侧重于提升整体的多样性与保真度，但对于生成内容或风格的细粒度控制能力有待加强。未来可以研究如何在形状空间或其他流形表示中结合属性解耦技术，以实现对生成图像特定属性，如姿态、表情、纹理细节的更精确控制的更精确、独立的控制。此外，开发显式的机制来评估和控制通过流形上增强生成的特征或图像的质量，避免在数据增强应用中引

入低质量样本干扰下游任务，也是一个关键且实际的研究问题。例如，在风格迁移应用中，探索如何通过控制在测地路径或曲面上的插值参数来实现对风格化程度的连续、平滑调节，将是提升用户体验的重要方向。

参考文献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[2] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International conference on machine learning. [S.l.]: PMLR, 2015: 2256-2265.

[3] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[4] KARRAS T, AITTALA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[J]. Advances in neural information processing systems, 2020, 33: 12104-12114.

[5] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 10684-10695.

[6] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of big data, 2019, 6(1): 1-48.

[7] OJHA U, LI Y, LU J, et al. Few-shot image generation via cross-domain correspondence[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 10743-10752.

[8] XIAO J, LI L, WANG C, et al. Few shot generative model adaption via relaxed spatial structural alignment[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 11204-11213.

[9] ZHAO S, LIU Z, LIN J, et al. Differentiable augmentation for data-efficient gan training[J]. Advances in neural information processing systems, 2020, 33: 7559-7570.

[10] VERMA V, LAMB A, BECKHAM C, et al. Manifold mixup: Better representations by interpolating hidden states[C]//International conference on machine learning. [S.l.]: PMLR, 2019: 6438-6447.

[11] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. [S.l.]: PMLR, 2021: 8748-8763.

[12] PATASHNIK O, WU Z, SHECHTMAN E, et al. Styleclip: Text-driven manipulation of stylegan imagery[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 2085-2094.

[13] TENENBAUM J B, SILVA V D, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. science, 2000, 290(5500): 2319-2323.

[14] BRONSTEIN M M, BRUNA J, LECUN Y, et al. Geometric deep learning: going beyond euclidean data[J]. IEEE Signal Processing Magazine, 2017, 34(4): 18-42.

[15] VAN DER MAATEN L, HINTON G. Visualizing data using t-sne.[J]. Journal of machine learning research, 2008, 9(11).

[16] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. [S.l.]: PmLR, 2020: 1597-1607.

[17] ZHU J Y, KRÄHENBÜHL P, SHECHTMAN E, et al. Generative visual manipulation on the natural image manifold[C]//European Conference on Computer Vision. [S.l.: s.n.], 2016: 597-613.

[18] BELKIN M, NIYOGI P, SINDHWANI V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.[J]. Journal of machine learning research, 2006, 7(11).

[19] KENDALL D G. Shape manifolds, procrustean metrics, and complex projective spaces[J]. Bulletin of the London mathematical society, 1984, 16(2): 81-121.

[20] KILIAN M, MITRA N J, POTTMANN H. Geometric modeling in shape space[J]. ACM Transactions on Graphics (SIGGRAPH), 2007, 26(3): #64, 1-8.

[21] HAN Y, WANG B, IDESAWA M, et al. Recognition of multiple configurations of objects with limited data[J]. Pattern Recognition, 2010, 43(4): 1467-1475.

[22] HAN Y, KOIKE H, IDESAWA M. Recognizing objects with multiple configurations[J]. Pattern Analysis and Applications, 2014, 17: 195-209.

[23] PASKIN M, BAUM D, DEAN M N, et al. A kendall shape space approach to 3d shape estimation from 2d landmarks[C]//European Conference on Computer Vision. [S.l.]: Springer, 2022: 363-379.

[24] FRIJI R, DRIRA H, CHAIEB F, et al. Geometric deep neural network using rigid and non-rigid transformations for human action recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 12611-12620.

[25] VADGAMA S, TOMCZAK J M, BEKKERS E J. Kendall shape-vae: Learning shapes in a generative framework[C]//NeurIPS 2022 Workshop on Symmetry and Geometry in Neural Representations. [S.l.: s.n.], 2022.

[26] HAN Y, WAN G, WANG B. Fagc:feature augmentation on geodesic curve in the pre-shape space[Z]. [S.l.: s.n.], 2023.

[27] DEVRIES T, TAYLOR G W. Dataset augmentation in feature space[Z]. [S.l.: s.n.], 2017.

[28] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: Beyond empirical risk minimization[C]//International Conference on Learning Representations. [S.l.: s.n.], 2018.

[29] LI P, LI D, LI W, et al. A simple feature augmentation for domain generalization[C]// Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 8886-8895.

[30] KUO C W, MA C Y, HUANG J B, et al. Featmatch: Feature-based augmentation for semi-supervised learning[C]//European Conference on Computer Vision. [S.l.: s.n.], 2020: 479-495.

[31] MANGLA P, KUMARI N, SINHA A, et al. Charting the right manifold: Manifold mixup for few-shot learning[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. [S.l.: s.n.], 2020: 2218-2227.

[32] KHAN A, FRAZ K. Post-training iterative hierarchical data augmentation for deep networks[J]. Advances in Neural Information Processing Systems, 2020, 33: 689-699.

[33] ZHOU K, YANG Y, QIAO Y, et al. Mixstyle neural networks for domain generalization and adaptation[J]. International Journal of Computer Vision, 2024, 132(3): 822-836.

[34] CHU P, BIAN X, LIU S, et al. Feature space augmentation for long-tailed data[C]// European Conference on Computer Vision. [S.l.: s.n.], 2020: 694-710.

[35] LIU D, ZHONG S, LIN L, et al. Feature-level smote: Augmenting fault samples in learnable feature space for imbalanced fault diagnosis of gas turbines[J]. Expert Systems with Applications, 2024, 238: 122023.

[36] KINGMA D P, WELLING M. Auto-encoding variational bayes[Z]. [S.l.: s.n.], 2022.

[37] HAN Y, LIU Y, CHEN Q. Data augmentation in material images using the improved hp-vae-gan[J]. Computational Materials Science, 2023, 226: 112250.

[38] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. ACM computing surveys (csur), 2020, 53(3): 1-34.

[39] ZHAO Y, DING H, HUANG H, et al. A closer look at few-shot image generation[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 9140-9150.

[40] WANG Y, GONZALEZ-GARCIA A, BERGA D, et al. Minegan: effective knowledge transfer from gans to target domains with few images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 9332-9341.

[41] SEO J, KANG J S, PARK G M. Lfs-gan: Lifelong few-shot image generation[C]// Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2023: 11356-11366.

[42] ZHAO Y, CHANDRASEGARAN K, ABDOLLAHZADEH M, et al. Few-shot image generation via adaptation-aware kernel modulation[J]. Advances in Neural Information Processing Systems, 2022, 35: 19427-19440.

[43] LIU B, ZHU Y, SONG K, et al. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[44] WANG Z, JIANG Y, ZHENG H, et al. Patch diffusion: Faster and more data-efficient training of diffusion models[J]. Advances in neural information processing systems, 2023, 36: 72137-72154.

[45] SHAHAM T R, DEKEL T, MICHAELI T. Singan: Learning a generative model from a single natural image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2019: 4570-4580.

[46] HINZ T, FISHER M, WANG O, et al. Improved techniques for training single-image gans[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.: s.n.], 2021: 1300-1309.

[47] WANG W, BAO J, ZHOU W, et al. Sindiffusion: Learning a diffusion model from a single natural image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[48] GUR S, BENAIM S, WOLF L. Hierarchical patch vae-gan: Generating diverse videos from a single sample[J]. Advances in Neural Information Processing Systems, 2020, 33: 16761-16772.

[49] KONG C, KIM J, HAN D, et al. Few-shot image generation with mixup-based distance learning[C]//European conference on computer vision. [S.l.]: Springer, 2022: 563-580.

[50] PORTILLA J, SIMONCELLI E P. A parametric texture model based on joint statistics of complex wavelet coefficients[J]. International journal of computer vision, 2000, 40: 49-70.

[51] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 2223-2232.

[52] PARK T, EFROS A A, ZHANG R, et al. Contrastive learning for unpaired image-to-image translation[C]//European Conference on Computer Vision. [S.l.: s.n.], 2020: 319-345.

[53] LIU R, WANG T, LI H, et al. Tmm-nets: transferred multi-to mono-modal generation for lupus retinopathy diagnosis[J]. IEEE Transactions on Medical Imaging, 2022, 42 (4): 1083-1094.

[54] ZHAO W, ZHU J, HUANG J, et al. Gan-based multi-decomposition photo cartoonization[J]. Computer Animation and Virtual Worlds, 2024, 35(3): e2248.

[55] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 1501-1510.

[56] LIU S, LIN T, HE D, et al. Adaattn: Revisit attention mechanism in arbitrary neural style transfer[C]//Proceedings of the IEEE/CVF international conference on computer vision. [S.l.: s.n.], 2021: 6649-6658.

[57] YU X, ZHOU G. Arbitrary style transfer via content consistency and style consistency[J]. The Visual Computer, 2024, 40(3): 1369-1382.

[58] ZHANG Y, TANG F, DONG W, et al. Domain enhanced arbitrary image style transfer via contrastive learning[C]//ACM SIGGRAPH 2022 conference proceedings. [S.l.: s.n.], 2022: 1-8.

[59] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[60] DENG Y, TANG F, DONG W, et al. Stytr2: Image style transfer with transformers[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 11326-11336.

[61] SU X, SONG J, MENG C, et al. Dual diffusion implicit bridges for image-to-image translation[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2023.

[62] WANG Z, ZHAO L, XING W. Stylediffusion: Controllable disentangled style transfer via diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2023: 7677-7689.

[63] ZHANG Y, HUANG N, TANG F, et al. Inversion-based style transfer with diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2023: 10146-10156.

[64] CHO H, LEE J, CHANG S, et al. One-shot structure-aware stylized image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2024: 8302-8311.

[65] TAN W R, CHAN C S, AGUIRRE H E, et al. Improved artgan for conditional synthesis of natural image and artwork[J]. IEEE Transactions on Image Processing, 2018, 28(1): 394-409.

[66] LI J, LI D, XIONG C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International conference on machine learning. [S.l.]: PMLR, 2022: 12888-12900.

[67] RUIZ N, LI Y, JAMPANI V, et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2023: 22500-22510.

[68] HU E J, YELONG SHEN, WALLIS P, et al. LoRA: Low-rank adaptation of large language models[C]//International Conference on Learning Representations. [S.l.: s.n.], 2022.

[69] GAL R, PATASHNIK O, MARON H, et al. Stylegan-nada: Clip-guided domain adaptation of image generators[J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 1-13.

[70] CROWSON K, BIDERMAN S, KORNIS D, et al. Vqgan-clip: Open domain image generation and editing with natural language guidance[C]//European Conference on Computer Vision. [S.l.]: Springer, 2022: 88-105.

[71] KWON G, YE J C. Diffusion-based image translation using disentangled style and content representation[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2023.

[72] PODELL D, ENGLISH Z, LACEY K, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis[C]//The Twelfth International Conference on Learning Representations. [S.l.: s.n.], 2024.

[73] YANG S, HWANG H, YE J C. Zero-shot contrastive loss for text-guided diffusion image style transfer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2023: 22873-22882.

[74] HE F, LI G, SUN F, et al. Freestyle: Free lunch for text-guided style transfer using diffusion models[Z]. [S.l.: s.n.], 2024.

[75] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross-attention control[C]//The Eleventh International Conference on Learning Representations. [S.l.: s.n.], 2023.

[76] TUMANYAN N, GEYER M, BAGON S, et al. Plug-and-play diffusion features for text-driven image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2023: 1921-1930.

[77] DONG W, XUE S, DUAN X, et al. Prompt tuning inversion for text-driven image editing using diffusion models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.: s.n.], 2023: 7430-7440.

[78] XU S, HUANG Y, PAN J, et al. Inversion-free image editing with language-guided diffusion models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2024: 9452-9461.

[79] JU X, ZENG A, BIAN Y, et al. Pnp inversion: Boosting diffusion-based editing with 3 lines of code[C]//The Twelfth International Conference on Learning Representations. [S.l.: s.n.], 2024.

[80] FLETCHER P T, LU C, PIZER S M, et al. Principal geodesic analysis for the study of nonlinear statistics of shape[J]. IEEE transactions on medical imaging, 2004, 23 (8): 995-1005.

[81] DRYDEN I L, MARDIA K V. Statistical shape analysis: with applications in r[M]. [S.l.]: John Wiley & Sons, 2016.

[82] PENNEC X. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements[J]. Journal of Mathematical Imaging and Vision, 2006, 25: 127-154.

[83] PENNEC X. Barycentric subspace analysis on manifolds[J]. The Annals of Statistics, 2018, 46(6A): 2711-2746.

[84] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[85] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//International conference on machine learning. [S.l.]: PMLR, 2017: 214-223.

[86] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of wasserstein gans[J]. Advances in neural information processing systems, 2017, 30.

[87] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[C]//International Conference on Learning Representations. [S.l.: s.n.], 2018.

[88] BROCK A, DONAHUE J, SIMONYAN K. Large scale GAN training for high fidelity natural image synthesis[C]//International Conference on Learning Representations. [S.l.: s.n.], 2019.

[89] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 4401-4410.

[90] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020: 8110-8119.

[91] KARRAS T, AITTALA M, LAINE S, et al. Alias-free generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2021, 34: 852-863.

[92] SAUER A, SCHWARZ K, GEIGER A. Stylegan-xl: Scaling stylegan to large diverse datasets[C]//ACM SIGGRAPH 2022 conference proceedings. [S.l.: s.n.], 2022: 1-10.

[93] HUANG Y L, YUAN X F. Styleterrain: A novel disentangled generative model for controllable high-quality procedural terrain generation[J]. Computers & Graphics, 2023, 116: 373-382.

[94] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2016: 770-778.

[95] TAKIDA Y, IMAIZUMI M, SHIBUYA T, et al. SAN: Inducing metrizability of GAN with discriminative normalized linear layer[C]//The Twelfth International Conference on Learning Representations. [S.l.: s.n.], 2024.

[96] BAO J, CHEN D, WEN F, et al. Cvae-gan: fine-grained image generation through asymmetric training[C]//Proceedings of the IEEE international conference on computer vision. [S.l.: s.n.], 2017: 2745-2754.

[97] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[98] JIANG Y, CHANG S, WANG Z. Transgan: Two pure transformers can make one strong gan, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.

[99] ESSER P, ROMBACH R, OMMER B. Taming transformers for high-resolution image synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2021: 12873-12883.

[100] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[C]//International Conference on Learning Representations. [S.l.: s.n.], 2021.

[101] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents[Z]. [S.l.: s.n.], 2022.

[102] LIAN L, LI B, YALA A, et al. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models[J]. Transactions on Machine Learning Research, 2024.

[103] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. [S.l.]: Springer, 2015: 234-241.

[104] DHARIWAL P, NICHOL A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.

[105] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[106] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018: 586-595.

[107] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13 (4): 600-612.

[108] HESSEL J, HOLTZMAN A, FORBES M, et al. Clipscore: A reference-free evaluation metric for image captioning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 7514-7528.

[109] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 2019, 32: 19427-19440.

[110] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[111] LIU S, ZHANG X, WANGNI J, et al. Normalized diversification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 10306-10315.

[112] MAO Q, LEE H Y, TSENG H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2019: 1429-1437.

[113] BENAIM S, WOLF L. One-sided unsupervised domain mapping[J]. Advances in neural information processing systems, 2017, 30.

[114] YANIV J, NEWMAN Y, SHAMIR A. The face of art: landmark detection and geometric style in portraits[J]. ACM Transactions on graphics (TOG), 2019, 38(4): 1-15.

[115] SI Z, ZHU S C. Learning hybrid image templates (hit) by information projection[J]. IEEE Transactions on pattern analysis and machine intelligence, 2011, 34(7): 1354-1367.

[116] WANG X, TANG X. Face photo-sketch synthesis and recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(11): 1955-1967.

[117] LIU Z, LUO P, WANG X, et al. Large-scale celebfaces attributes (celeba) dataset[J]. Retrieved August, 2018, 15(2018): 11.

[118] DECOST B L, HECHT M D, FRANCIS T, et al. Uhcsdb: ultrahigh carbon steel micrograph database: tools for exploring large heterogeneous microstructure datasets[J]. Integrating Materials and Manufacturing Innovation, 2017, 6: 197-205.

[119] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. [S.l.]: PMLR, 2019: 6105-6114.

[120] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2017: 4700-4708.

[121] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2022: 11976-11986.

[122] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]//International conference on machine learning. [S.l.]: PMLR, 2021: 10347-10357.

[123] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. [S.l.: s.n.], 2016: 785-794.

[124] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[C]// International Conference on Learning Representations. [S.l.: s.n.], 2021.

[125] KWON G, YE J C. Clipstyler: Image style transfer with a single text condition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2022: 18062-18071.

[126] BARANCHUK D, VOYNOV A, RUBACHEV I, et al. Label-efficient semantic segmentation with diffusion models[C]//International Conference on Learning Representations. [S.l.: s.n.], 2022.

[127] ESSER P, KULAL S, BLATTMANN A, et al. Scaling rectified flow transformers for high-resolution image synthesis[C]//International Conference on Machine Learning. [S.l.]: PMLR, 2024: 12606-12633.

[128] LABS B F. Flux[EB/OL]. 2024. https://github.com/black-forest-labs/flux.

[129] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. [S.l.]: Ieee, 2009: 248-255.

[130] HORE A, ZIOU D. Image quality metrics: Psnr vs. ssim[C]//2010 20th international conference on pattern recognition. [S.l.]: IEEE, 2010: 2366-2369.

[131] VON PLATEN P, PATIL S, LOZHKOV A, et al. Diffusers: State-of-the-art diffusion models[J/OL]. GitHub repository, 2022. https://github.com/huggingface/diffusers.

攻读硕士学位期间取得的研究成果

一、主要作者研究成果

[1] Han Y, Ruan L, Wang B. Few-shot image generation via information transfer from the built geodesic surface. 已投稿至 Pattern Recognition (中科院一区 Top 期刊)，目前状态为大修已返回 (Major Revision)，(预印本: arXiv:2401.01749, 2024)。(导师一作，本人二作)

[2] Han Y, Ruan L, Wang B. Geodesic Feature Augmentation for Zero-shot Text-Guided Diffusion Style Transfer. 已投稿至 The Visual Computer (CCF-C 类期刊)，目前状态为大修已返回 (Major Revision)，(预印本: arXiv:2408.10533, 2024)。(导师一作，本人二作)

[3] Han Y, Ruan L, Wang B. Feature Augmentation via Dirichlet Mixup for Text-guided Diffusion Image Style Transfer. 已被 2025 6th International Conference on Computer Information and Big Data Applications (CIBDA 2025) (EI 检索会议) 录用，待发表。(导师一作，本人二作)

二、参与的论文修订与发表工作

在以下已发表论文的工作中，本人主要负责了审稿意见的回复与论文修改工作：

[1] Han, Y., Li, R., Wang, B., Ruan, L., & Chen, Q. (2024). A pseudo-labeling based weakly supervised segmentation method for few-shot texture images. Expert Systems with Applications (中科院一区 Top 期刊), 238, 122110.

[2] Chen, Q., Wei, H., Wang, B., Ruan, L., & Han, Y. (2023). Material structure segmentation method based on graph attention. Materials Today Communications, 35, 105941.

[3] Han Y, Li R, Ruan L, et al. (2024). Statistics and Analysis of Lath Martensite Transformation Based on In-Situ Observation and Video Processing. Physics of Metals and Metallography, 125(Suppl 1): S106-S120.

三、专利

[1] 专利名称：一种基于形状空间理论的图像生成方法，发明人：韩越兴、阮礼恒、王冰。申请号或专利号：2023117418451，申请公布号：CN117746178A，申请日：

2023 年 12 月 18 日，公开日：2024 年 03 月 22 日。(导师一作，本人二作)

四、软件著作权

[1] 软件名称：融合深度学习方法和形状空间理论的小样本图像生成平台软件 V1.0，开发人：韩越兴、阮礼恒、王冰。登记号：2023SR1425482，申请人：上海大学，开发完成日期：2023 年 6 月 30 日，登记日期：2023 年 7 月 30 日。

[2] 软件名称：融合深度学习方法和形状空间理论的文本指导图像风格迁移平台软件 V1.0，开发人：韩越兴、阮礼恒、王冰。登记号：2024R11L1524675，申请人：上海大学，开发完成日期：2024 年 6 月 2 日，登记日期：2024 年 7 月 1 日。

致　谢

七年的上大时光即将画上句号。回望这段旅程，从初入校园到完成学业，离不开师长、家人和朋友们的支持与鼓励。值此之际，谨向所有帮助过我的人们致以最诚挚的谢意。

首先，我要诚挚地感谢我的导师韩越兴教授。本研究从最初的构想到最终的成果，都凝聚了韩老师的大量心血。在学术探索中，韩老师对我影响至深，他精益求精的治学态度和对学术细节的极致追求，为我树立了做学问的标杆。生活上，韩老师风趣的谈吐和宽和的为人，也极大缓解了科研的压力，营造了融洽活跃的团队氛围。这段师生之谊，是我求学生涯中最宝贵的收获之一。同时，感谢陈侨川老师在科研思路上提出的宝贵建议，总能带给我新的启发。也感谢张瑞老师和孙妍老师在组会中的精辟见解，有效拓宽了我的学术视野。祝愿各位老师工作顺利，生活愉快。

其次，我要由衷地感谢我的家人。感谢父母的养育之恩，你们为我营造了幸福和睦的家庭环境，让我能快乐地成长，几乎未曾经历大的风浪。你们的理解与支持，是我安心求学、勇于探索的最大动力。也感谢我的姐姐、表姐以及表兄弟们，你们日常的关心、及时的帮助以及作为榜样给我带来的力量，都让我受益良多。

实验室的时光同样丰富而难忘。感谢所有同门伙伴，是你们让这里充满了家一般的归属感。感谢师兄师姐们在科研、工作和生活上分享的宝贵经验，让我少走了很多弯路。也感谢充满活力的师弟师妹们，你们的勤奋努力，也时常激励着我。特别感谢我的同届朋友们，我们不仅在科研上互勉互助，更在生活中建立了深厚的友谊。忘不了一起游泳健身的酣畅，也忘不了一同在实验室奋斗的夜晚——正是这些真实的瞬间，让略显单调的科研生活变得有滋有味，充满温暖。拥有这样真挚的同窗情谊，我倍感幸运，也将永远珍惜。

此外，我要将最特别的感谢给予我在读研过程中遇到的另一半，也是我最重要的朋友。与你相处是一种奇妙的“共处式充电”，有时仅是安静地待在彼此身边，那份不言而喻的默契与安心，就能消解我所有的疲惫与烦恼。你的陪伴是我重要的精神支柱，和你共度的时光，是我汲取能量、重获勇气的源泉。感谢你的出现，也期盼未来的道路能继续与你并肩同行。

最后，感谢上海大学。从本科到研究生，七年的时光承载了我的青春，也见证了我从学生到准社会人的转变。校园里的一草一木，都记录着这段宝贵的成长印记。感谢这段旅程赋予我的学识、友谊与成长，也感谢始终坚持的自己。

行文至此，心中唯有感恩。愿我们在未来的道路上，都能行稳致远，前程似锦。

阮礼恒

上海大学

2025 年 5 月 8 日