基于神经网络的图像风格迁移算法综述

王 伟,张静宜,温玉辉,魏云超

(北京交通大学计算科学与技术学院,北京 100044)

摘 要: 风格迁移作为图像编辑领域的一个关键研究方向,在艺术创作等领域展现出广泛的应用前景. 自 Gatys 等人提出使用深度卷积特征间相关性捕获纹理信息并基于此实现风格迁移后,大量基于神经网络的风格迁移算法不断涌现. 近年来随着各式生成模型的兴起,将生成对抗网络、扩散模型等生成模型引入风格迁移工作获得了新的关注. 此外,图像-文本跨模态任务的突破使得文本引导条件下的图像风格迁移成为可能. 本文对当前先进的研究方法进行分类和描述. 具体地,依据引导条件差异,将现有方法划分为图像引导的图像风格迁移方法、文本引导的图像风格迁移方法;依据网络架构的不同,将现有方法细分为基于自编码器的方法、基于生成对抗网络的方法、基于扩散模型的方法以及基于其他模型架构的方法,对当前图像风格迁移技术的研究进行全面的综述与分析. 随后,介绍了图像风格迁移任务的数据集和评价体系,并从定量与定性方面对部分最先进的图像风格迁移方法进行实验和比较. 最后,讨论了当前图像风格迁移技术面临的挑战,并对未来研究方向提出了展望.

关键词: 图像风格迁移;神经网络;图像编辑;多模态任务;计算机视觉;深度学习

基金项目: 中央高校基本科研基金(No.2022XKRC015);国家自然科学基金(No.62372033)

中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2025)05-1692-21

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240930

Neural Network Based Image Style Transfer: A Survey

WANG Wei, ZHANG Jing-yi, WEN Yu-hui, WEI Yun-chao

(*School of Computer Science and Technology*, *Beijing Jiaotong University*, *Beijing* 100044, *China*)

Abstract: As a key research direction in the field of image editing, style transfer has shown a broad applications in artistic creation and related fields. Since Gatys et al. proposed the use of deep convolutional inter-feature correlations to capture texture information for style transfer, numerous neural style transfer algorithms have emerged. Recently, with the rise of various generative models, particularly the introduction of generative adversarial networks and diffusion models, style transfer work has gained new attention. Additionally, breakthroughs in image-text cross-modal tasks have made text-guided image style transfer possible. This paper presents a comprehensive review of the latest developments in style transfer techniques, classifying methods into image-guided and text-guided categories based on the guiding conditions. Furthermore, the methods are categorized into autoencoder-based approaches, GAN-based methods, diffusion model-based methods, and other architectural variants. This paper also introduces relevant dataset and evaluation metrics for image style transfer tasks, and compares state-of-the-art methods in terms of quantitative and qualitative aspects. Finally, the paper discusses the challenges and g provides insights into potential future research directions.

Key words: image style transfer; neural network; image editing; multimodal tasks; computer vision; deep learning

Foundation Item(s): Fundamental Research Funds for the Central Universities (No.2022XKRC015); National Natural Science Foundation of China (No.62372033)

1 引言

随着计算机视觉和深度学习技术的飞速发展,图像风格迁移已成为近年来图像处理与生成领域的一个热门研究方向. 风格迁移技术旨在根据参照将一种艺术风格或视觉特征转移到另一个目标图像上,从而创造出新颖且富有表现力的视觉作品. 风格迁移技术具有广泛的应用前景,在艺术创作与设计方面,可以使用

收稿日期:2024-10-16;修回日期:2025-04-21;责任编辑:宋甲英

图像风格迁移进行风格融合或转换,进行电影特效制作、虚拟现实技术合成,或者应用到设计中,快速获得设计灵感和效果. 在工程领域,通过风格迁移方法将图像从源域转换至目标风格域,可以快速实现图像处理. 例如通过风格迁移实现水下光学图像到合成孔径声纳图像的快速处理[1],或者使用风格转换不同源传感器的遥感图像[2],为海洋勘探、水下地形测绘等工程应用提供极大的便捷性. 此外,风格迁移技术还展现出作为数据增强手段的巨大潜力. 文献[3,4]通过风格混合策略增加训练数据的多样性,进而提升语义分割的性能;文献[5]利用源域风格对目标数据进行风格化处理,实现无监督的域自适应. 这些应用实例充分证明,风格迁移不仅是一种创新的图像处理技术,同时也是一种高效且实用的数据增强方法.

在计算机视觉发展的早期,传统的图像风格迁移方法直接使用滤波器对内容图像进行处理[6,7],或者应用机器学方法进行图片风格融合[8]. 但这些传统非参数的风格迁移方法只能通过提取图像的低层次特征(色彩、纹理等)来进行纹理合成,没有充分挖掘高级图像结构及语义特征. 随着深度学习的兴起,Gatys等人[9]证明了深度卷积特征矩阵之间的相关性可以很好地捕获纹理信息,并据此提出使用卷积神经网络(Convolutional Neural Networks,CNN)迭代优化来实现风格迁移任务[10]. 这一开创性成果推动了基于神经网络的图像风格迁移方法的发展,这种方法在艺术创作中得到了广泛应用.

随着自然语言处理技术的进步,文本引导的图像风格迁移技术逐渐成为新的研究热点. 以CLIP(Contrastive Language-Image Pretraining)[11]为代表的跨模态模型展现了出色的性能,实现了图像与文本在特征空间的对齐,大力推动了基于文本进行图像风格迁移任务的发展. 文本引导的图像风格迁移算法通过理解文本描述中的语义信息,指导图像生成过程. 从早期研究[12~14]中仅依赖于粗略的风格描述(如“素描”“油画”)作为引导,到近年来一些研究[15~17]依据详尽的文本描述,更精确地控制生成高度个性化的图像风格. 此外,部分前沿工作[18,19]同时支持将图像与文本作为多模态输入条件,进一步扩展了任务边界. 这种跨模态条件引导的风格迁移技术为图像生成提供了更多元的控制手段,使得用户能够根据自己的想象和描述创造出个性化的图像作品.

本文对图像风格迁移领域的研究进展进行了系统的分析与总结. 虽然已有部分文献综述了图像风格迁移相关的算法[20,21],但这些综述对算法的分类和分析仍不够细致. 例如,文献[20]聚焦于图像引导的风格迁移算法,文献[21]则主要回顾了基于生成对抗网络(Generative Adversarial Network,GAN)[22]的风格迁移方法,而对基于扩散模型[23]的方法探讨不足,同时对不同风格迁移方法的优势和差异性缺乏关注. 本文从引导条件和网络架构两个角度对现有的图像风格迁移方法进行了更为系统的分类与总结,同时详细介绍了数据集和评价指标. 通过对这些技术的梳理和分析,旨在为读者提供一个全面的视角,以了解风格迁移技术的最新发展和未来趋势.

2　问题与挑战

图像风格迁移算法旨在将参照风格的视觉特征与目标内容相结合,以实现图像风格的多样化与个性化. 具体来说,图像引导的风格迁移通常利用一幅或多幅风格参考图像来指导目标图像的风格转换;而文本引导的风格迁移则通过解析文本描述中的风格概念,将其转化为具体的视觉风格并应用于图像. 这一过程不仅要求算法准确捕捉源参照(如图像或文本描述)中的风格信息,还需要将其有效地转换并应用于目标图像中,同时保持图像内容的完整性与真实性. 虽然图像风格迁移具有广泛的应用场景,但它是一个具有挑战性的研究任务,主要体现在如下方面.

(1)风格的抽象性与多样性. 在图像引导的风格迁移任务中,风格的表现形式多种多样,可以是油画风格、水彩画风格等各种形式的艺术风格,也可以是具有特定结构和纹理的视觉外观,甚至对于色彩的偏好也将构成图像的风格特质. 而在文本引导的风格迁移任务中,文本风格描述往往缺乏明确的视觉特征界定,如“梦幻般的”或“复古风格”,这些模糊且富有想象力的词汇要求算法具备从抽象概念中提取具体视觉元素的能力,从而实现风格的有效迁移.

(2)跨模态的语义理解与转换. 针对文本引导条件下的风格迁移任务,需要跨越图像与文本之间的语义鸿沟,深入理解文本中描述的风格概念. 此外,还需要考虑如何设计风格融合机制,以确保风格化结果的准确性和自然性.

(3)内容结构的完整性. 在风格迁移的过程中,保持图像内容的完整性和真实性同样至关重要. 如果通过不恰当的方式引入新风格,可能会造成图像内容的扭曲或破坏.

3　基于不同引导条件的风格迁移

图1按照时间顺序梳理了本文所综述的主要算法,纵轴根据引导条件划分为两个部分,上半部分为图像引导的风格迁移,下半部分为文本引导的风格迁移,对应第3节内容;同时使用不同符号对模型架构进行区分,包括自编码器、GAN、扩散模型和其他架构,对应第4节内容.

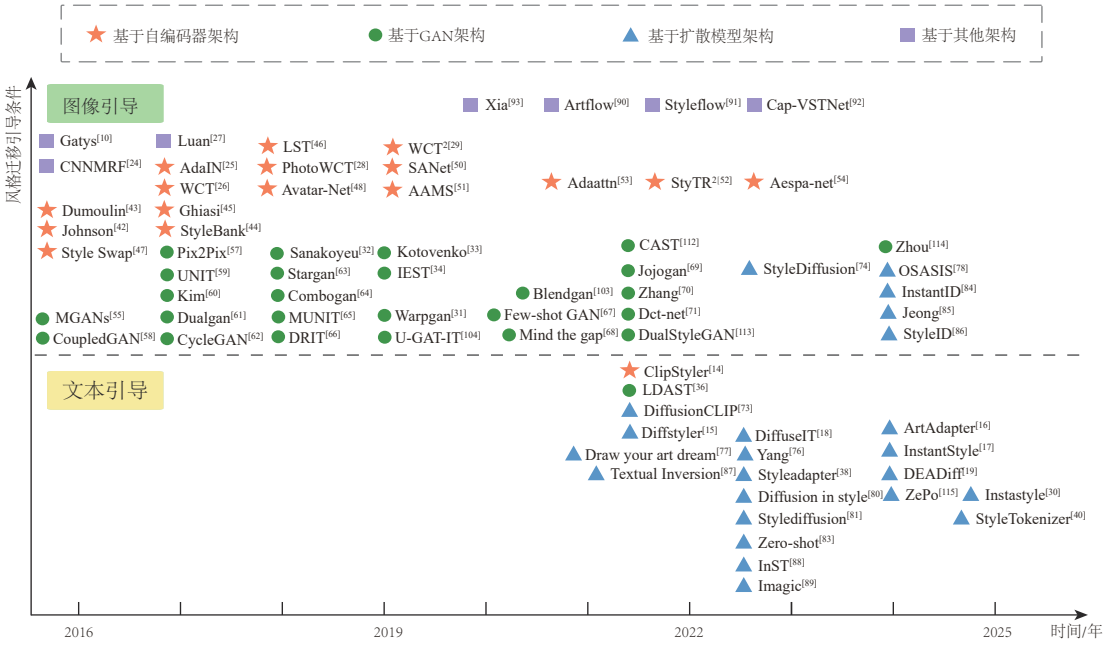


图1 风格迁移算法时间轴

从图1中可以看出，自Gatys等人[10]首次将神经网络引入风格迁移任务以来，该领域的研究在较长一段时间内主要围绕自编码器和生成对抗网络架构展开，且主要聚焦于图像引导的风格迁移算法. 然而，在实际生活应用中，用户期望的风格可能不再局限于一张图像，而是涵盖更加广泛、多元的媒介与表达方式. 为应对这一趋势，风格迁移算法的研究经历了从单一图像引导到跨模态、多模态引导条件的扩展. 这一转变的重要推力是2021年左右跨模态理解技术的突破性进展，特别是扩散模型在文本到图像生成方面所展现出的卓越能力. 自此，扩散模型逐渐成为风格迁移算法研究的主流方向，文本引导的图像风格迁移也受到越来越多的关注. 基于此，本节将聚焦于引导条件，对图像引导与文本引导的风格迁移算法进行深入分析.

3.1 图像引导的风格迁移

图像引导的风格迁移方法往往以内容图像和风格图像作为输入，在不损害原内容图像的结构细节的情况下，将其风格转换为参考图像所蕴含的独特艺术风格（包括色调、纹理等）. 自然地，下面对其进行进一步细分介绍.

3.1.1 单一风格图像引导的方法

在基于神经网络的风格迁移算法被提出后的很长一段时间内，引导条件仅为单张风格图像. 2016年，Gatys等人[10]首次证实了基于神经网络实现风格迁移的可行性. 他们将图片作为变量，通过VGG（Visual Geometry Group）网络提取特征，计算内容和风格损失并不断迭代优化，从而实现风格融合. 其风格损失函数如式(1)所示：

$$\mathcal{L}_{\mathrm{s}}=\sum_{l=1}^{L}\frac{1}{4N_l^2M_l^2}\sum_{i,j}\left(\boldsymbol{G}_{ij}^{l}-\boldsymbol{S}_{ij}^{l}\right)^2 \tag{1}$$

其中，L表示选取的网络层数，将特征图展平为$N_l\times M_l$后计算格拉姆矩阵（Gram Matrix）$\boldsymbol{G}^l$和$\boldsymbol{S}^l$（如图2所示），最小化它们的差异即可实现风格对齐.

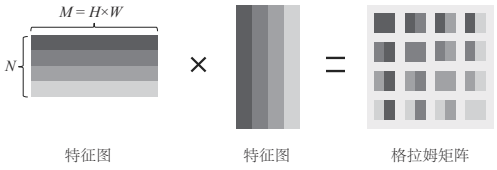


图2 格拉姆矩阵计算过程

这一方法成为风格迁移的标准，随后大量基于此框架的变体不断涌现，如结合马尔科夫随机场[24]和特征二阶统计量[25,26]的方式. 基于单张图像的风格迁移方法取得了快速发展，能够在较短时间内生成高质量的转换结果.

许多研究针对不同应用场景进一步扩展. 例如，照片级真实风格迁移[27~29]在不损害内容细节的前提下解决了结构伪影和失真问题；肖像风格化则在保持个人身份的基础上实现更具创造性的风格转换[30,31]，这些研究在艺术创作领域得到广泛的潜在应用.

3.1.2 风格图像集合引导的方法

部分工作[32,33]认为相较于单张图像，一组图像集合能够更好地表达艺术作品的风格. 以一些代表性工作为例，文献[32]通过从大型艺术数据集中选择与参

考图像相关的集合,基于风格感知损失实现了更精确的风格表达;文献[33]则结合对抗性损失实现特定内容的风格转换. Chen 等人[34]提出了基于对比学习的内部-外部风格迁移方法,以提升风格传递的质量.

3.2　文本引导的风格迁移

文本引导的风格迁移需要通过多模态模型实现不同模态信息的对齐,CLIP[11]使用图像编码器和文本编码器将图像和文本转换为特征表示,通过对比学习使图像和文本在特征空间中对齐. 其强大的跨模态理解能力在文本引导的图像风格迁移任务中得到了广泛应用. 举例来说,文献[12]将 CLIP 与 StyleGAN[35]相结合,基于图像空间变化和文本描述变化的方向共线性假设实现图像编辑. 受此启发,文献[13]以文本提示作为监督来实现生成器的域适应. 其中,为了在 CLIP 空间中优化风格迁移,提出了全局和局部方向性损失. 全局损失计算生成的图像 I_{cs} 与给定目标文本 t_s 之间的余弦距离,如下所示:

$$\mathcal{L}_{\text{global}} = D_{\text{CLIP}}\left(I_{cs}, t_s\right) \tag{2}$$

仅使用全局 CLIP 损失会破坏图像质量,因此引入了局部方向性损失,以对齐文本和图像对的嵌入之间的方向,其风格损失函数为 $\mathcal{L}_{\text{clip}} = 1 - \dfrac{\Delta \boldsymbol{I} \cdot \Delta \boldsymbol{T}}{|\Delta \boldsymbol{I}||\Delta \boldsymbol{T}|}$. 其中 $\Delta \boldsymbol{T} = \boldsymbol{E}_{\text{T}}\left(t_s\right) - \boldsymbol{E}_{\text{T}}\left(t_o\right)$, $\Delta \boldsymbol{I} = \boldsymbol{E}_{\text{I}}\left(I_{cs}\right) - \boldsymbol{E}_{\text{I}}\left(I_c\right)$, E_{T} 和 E_{I} 分别表示 CLIP 文本编码器和图像编码器,I_{cs} 和 I_c 分别为风格化图像和内容图像,t_s 和 t_o 分别是目标风格和输入内容描述,通过最小化损失函数,可以将文本间的差异转移到风格化图像中,过程如图 3 所示. 在 CLIP 空间的全局和局部方向性损失监督下,图像仅在不同风格的语义空间上发生变化,实现更多样性的文本驱动下的风格迁移. 文献[14]进一步设计了基于多视图增强图像块(patch)的 CLIP 方向损失,使风格纹理更加生动.

尽管这些技术在文本引导的风格迁移中取得了一定进展,但风格与语义的解耦仍是一个重要挑战. 文献[34]指出,当前基于 CLIP 的方法未能完全剥离风格中的语义干扰,因此通过自编码器(AutoEncoder, AE)[36]结合对比学习提升风格迁移效果,文献[19]则通过 Q-Former 模型解耦风格与语义表示. 然而,这些方法对数据集具有较高要求.

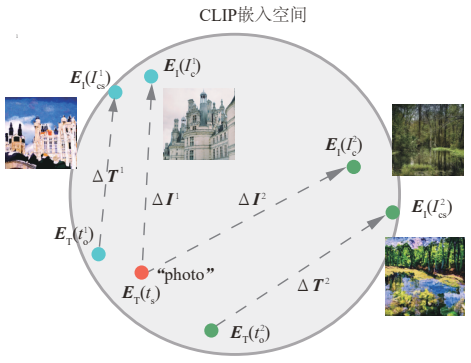


图 3　CLIP 方向损失

此外,随着文本到图像生成模型的兴起,一些方法利用其生成能力,以文本提示和风格参考图像为输入进行风格化图像生成. 此类方法可以结合 ControlNet[37]等控制手段保留参考内容图像的边缘结构,从而应用于图像风格迁移任务. 但风格参考图像中风格和语义信息的耦合问题可能导致风格参考和文本语义间的冲突,从而影响生成效果. 对此,文献[38]通过双路径交叉注意力模块处理文本提示和风格参考,减少了语义与风格的强耦合. Wang 等人[17]发现在 U-Net 中可以通过特定层实现风格和内容的解耦,提升风格迁移效果. 文献[39]发现风格化参考图像的反演噪声中携带风格信号,利用其实现了高质量的风格化生成. 文献[40]设计一种风格标记器将风格表示与文本表示对齐,以最大限度地减少参考风格对文本提示有效性的影响. 尽管当前的研究已取得一定进展,但仍未能有效解决输入条件间语义冲突带来的干扰问题. 因此,未来的研究还需深入探索跨模态信息的风格的解耦与对齐机制,以优化风格迁移效果并提高生成质量和效率.

3.3　小结

从风格迁移的引导方式角度出发,本节讨论了基于图像引导和文本引导的风格迁移方法. 传统的图像引导方法通过从视觉特征中提取风格,能够有效地转移色彩、纹理等低级特征,且基于图像集合的方式在捕捉细节(如笔触)方面表现更佳. 相比之下,文本引导方法通过语义描述提供了更为丰富的控制维度,能够灵活地调整几何形态上的风格,从而实现更加个性化和多样化的图像生成. 然而,由于跨模态模型的能力限制,现有方法可能无法准确地捕捉用户的意图,并有效传达所期望的风格信息. 此外,文本信息中风格与内容的纠缠仍是该领域亟待解决的关键问题.

4　基于不同网络架构的风格迁移

近年来,随着深度学习技术的飞速发展,特别是神经网络架构的多样化,图像风格迁移的方法也在不断演进. 本节基于模型架构的差异对其进行分类总结,如图 4 所示. 现有方法大致分为 4 类:基于自编码器的方法[41],利用其重构特性实现风格迁移;基于 GAN 的方法[22],通过生成器与判别器对抗训练实现风格转换;基于扩散模型的方法[23],通过去噪模拟数据生成风格;最后还包括了基于其他网络架构的风格迁移方法,展现了模型架构在风格迁移领域的多样性和创新性. 接下来将逐一介绍这些方法.

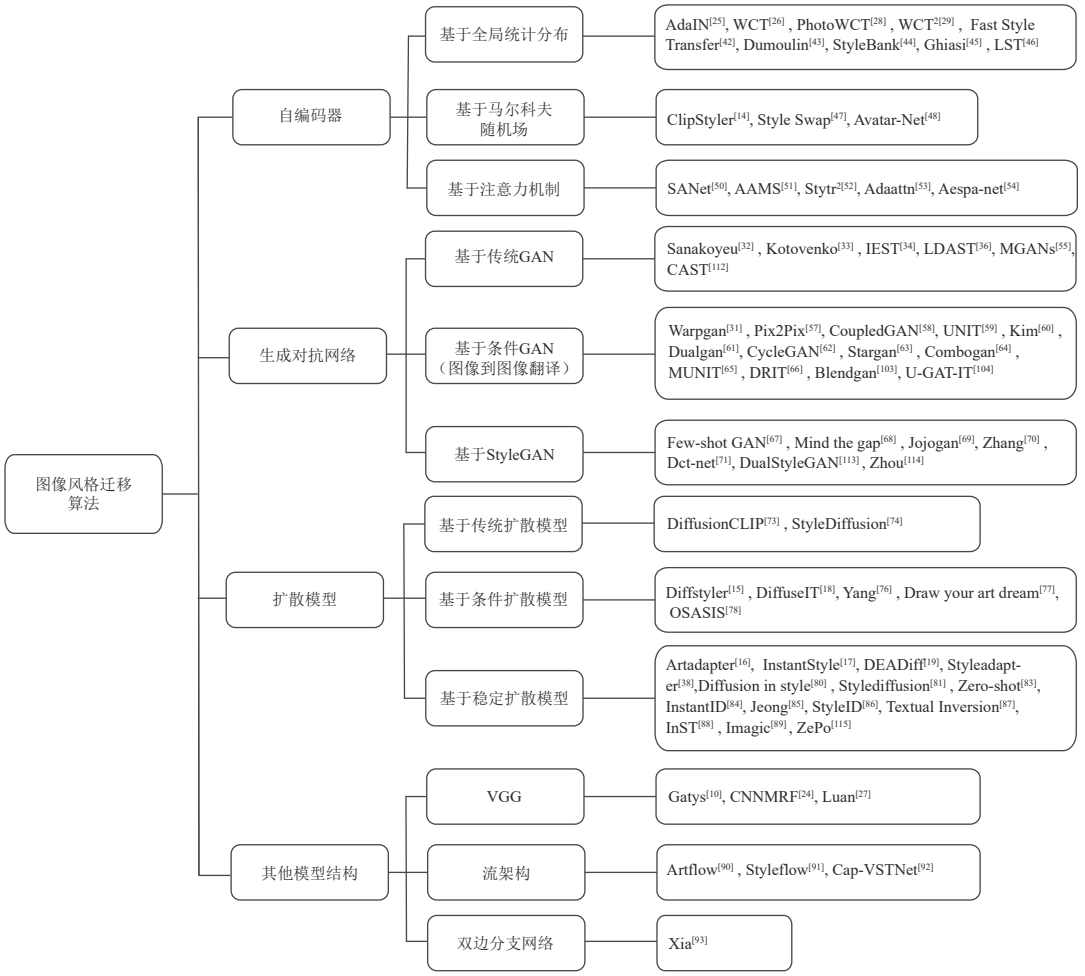


图4 基于网络架构进行分类的图像风格迁移算法

4.1 基于自编码器的风格迁移

自编码器由编码器(encoder)和解码器(decoder)两部分组成. 编码器将输入数据压缩为低维空间表征,解码器将该低维表征解码进行重新生成. 在图像重建的过程中,通过特定的训练策略和架构设计,可以实现风格注入. 如图4所示,基于自编码器结构的图像风格迁移方法主要通过基于全局统计分布、基于马尔科夫随机场和基于注意力机制3种手段实现.

4.1.1 基于全局统计分布实现风格迁移

Johnson等人[42]首次利用自编码器实现风格迁移,通过编码器、残差块和解码器构建风格变换网络,借助VGG-16计算全局内容和风格损失. 为提升灵活性,Dumoulin等人[43]提出了条件实例归一化(Conditional Instance Normalization,CIN),归一化内容特征后,利用风格参数缩放和平移. 而文献[44]设计了StyleBank模块编码风格信息. 这些方法将每种风格投影为特征参数,通过一次前馈过程即可实现多种风格化.

后续研究进一步提出了通用风格迁移方法. Ghiasi等人[45]通过风格预测网络生成任意风格参考的特征向量. 文献[25]提出自适应实例归一化(Adaptive Instance Normalization,AdaIN),通过对齐内容与风格特征的均值和方差进行风格迁移. 文献[26]则采用白化和着色变换(Whitening and Coloring Transform,WCT)实现快速风格转换. 尽管AdaIN和WCT实现了快速的风格迁移,但高维特征向量的二阶统计计算代价是昂贵的.

为简化计算,Li等人[46]将风格迁移简化为线性变换问题,其优化目标如式(3)所示. 其中,$\bar{F}_{\mathrm{c}}$表示内容特征,$\boldsymbol{T}$为变换矩阵,$\bar{F}_{\mathrm{d}}$表示风格变换后的特征,$\bar{F}$表示对$\boldsymbol{F}$零均值化处理,$\bar{\boldsymbol{\phi}}_{\mathrm{s}}$表示参考风格特征$\boldsymbol{F}_{\mathrm{s}}$的非线性映射. 当$\boldsymbol{\phi}_{\mathrm{s}}=\boldsymbol{F}_{\mathrm{s}}$时,$\boldsymbol{F}_{\mathrm{d}}\boldsymbol{F}_{\mathrm{d}}^{\mathrm{T}}$可以视为格拉姆矩阵,因此与式(1)建立联系. 通过学习变换矩阵,模型实现了通用的风格迁移.

$$\begin{aligned}\boldsymbol{F}_{\mathrm{d}}^{*}=\arg\min_{\boldsymbol{F}_{\mathrm{d}}}\quad & \frac{1}{NC}\left\|\bar{F}_{\mathrm{d}}\bar{F}_{\mathrm{d}}^{\top}-\bar{\boldsymbol{\phi}}_{\mathrm{s}}\bar{\boldsymbol{\phi}}_{\mathrm{s}}^{\top}\right\|_{\boldsymbol{F}}^{2} \\ \text{s.t.}\quad & \bar{F}_{\mathrm{d}}=\boldsymbol{T}\bar{F}_{\mathrm{c}}\end{aligned} \tag{3}$$

4.1.2 基于马尔科夫随机场实现风格迁移

尽管基于全局信息的风格迁移方法取得了一定效

果,但在处理复杂结构时,结果可能出现不正确的风格模式. 因此,部分研究采用基于patch的风格迁移算法. patch为局部k*k像素区域,通过马尔科夫随机场匹配最相关的局部特征进行风格融合. StyleSwap[47]使用编码器实现并行的patch匹配与风格融合,并训练一个逆向网络重建风格化结果. 该方法实现了通用风格迁移,但受到风格化效果弱的限制. 而Sheng等人[48]在AE结构中引入逐级风格增强,设计了一个具有多尺度风格迁移模块的沙漏网络. 然而,由于此类方法更侧重相似块之间的风格迁移,这可能导致伪影的产生和内容细节的丢失.

4.1.3　基于注意力机制实现风格迁移

注意力机制最早被用于自然语言处理领域,Transformer[49]使用注意力机制处理序列数据,这一创新显著增强了模型的表征能力和并行计算能力,也激发了视觉领域的一系列创新. 在注意力机制中,注意力图将某个位置的响应计算为所有位置的特征的加权总和,这有助于捕获图像语义结构信息,实现更精细的风格化,同时更好地保留图像的内容细节. Park等人[50]首次将自注意力机制引入风格迁移,提出一种风格注意力网络,根据内容图像的语义分布灵活融合局部风格. Yao等人[51]通过多尺度风格交换和多笔触融合模块处理不同层次的细节. Deng等人[52]通过感知位置编码捕捉长距离信息,解决内容泄露问题.

此外,一些研究结合全局统计特征和注意力机制实现风格迁移. Liu等人[53]提出加权统计量与AdaIN相结合的方法,Hong等人[54]则度量特征图案的可重复性,在注意力模块和全局特征统计转换模块间取得权衡,提升风格转换效果.

4.2　基于生成对抗网络的风格迁移

在本节中,对使用生成对抗网络架构进行图像风格迁移的方法进行总结与介绍,如图4所示,围绕传统GAN结构、条件GAN结构以及StyleGAN[35]展开探讨.

4.2.1　基于传统GAN实现风格迁移

GAN是一种生成模型,灵感源自博弈论中的二人零和博弈,让生成器与判别器在相互博弈中进行学习,训练目标为

$$\min_{\boldsymbol{G}}\ \max_{\boldsymbol{D}} \mathrm{E}_{x\sim p_{\mathrm{data}}}\left[\log \boldsymbol{D}(\boldsymbol{x})\right]+E_{z\sim p_z}\left[\log\left(1-\boldsymbol{D}\left(\boldsymbol{G}(z)\right)\right)\right] \tag{4}$$

其中,$\boldsymbol{x}$为真实样本,z为随机噪声,判别器$\boldsymbol{D}(\boldsymbol{x})$学习区分真实样本和生成样本,而生成器$\boldsymbol{G}$通过生成逼近真实分布的样本使判别器难以区分.

Li等人[55]基于此提出了马尔科夫生成对抗网络(Markovian GANs,MGANs),将风格迁移视为匹配源图像和目标风格分布的任务. 该模型使用判别器不断学习区分不恰当风格化的patch,同时训练一个变分自动编码器作为生成器,实现快速高效的风格迁移. 但这种方法缺乏对高级语义信息的理解,在处理非纹理风格迁移时效果较差.

4.2.2　基于条件GAN实现风格迁移

条件生成对抗网络(Conditional GAN,CGAN)[56]通过额外的引入条件信息提高了生成的可控性. 其目标函数为

$$\min_{\boldsymbol{G}}\ \max_{\boldsymbol{D}} \mathrm{E}_{y,x}\left[\log \boldsymbol{D}\left(\boldsymbol{y},\boldsymbol{x}\right)\right] +\mathrm{E}_{y,z}\left[\log\left(1-\boldsymbol{D}\left(\boldsymbol{y},\boldsymbol{G}\left(\boldsymbol{y},z\right)\right)\right)\right] \tag{5}$$

指导条件$\boldsymbol{y}$与随机噪声z共同输入生成器进行图像生成,可以理解为传统GAN学习的是噪声z到$\boldsymbol{x}$的映射,而cGAN学习的是从$\boldsymbol{y}$和z到$\boldsymbol{x}$的映射. 基于此,Isola等人[57]在2017年提出一种基于cGAN的通用框架Pix2Pix,解决包含风格迁移在内的图像到图像翻译任务. Pix2Pix将内容图像作为条件信息,使用成对的图像学习从输入到参考风格之间的映射. 训练完成后,将任意内容图像输入生成器中,即可获得指定的风格迁移结果.

然而,真实照片与其风格化结果的成对数据基本不存在,数据集的获取代价是昂贵的. 因此,许多研究工作[58-61]转而探索无监督的图像到图像翻译方法,以减轻对配对训练数据的严格依赖. 如Liu等人[58]提出的UNIT,假定处于不同域共享同一隐空间,如图5(a)所示,通过权重共享约束和对抗性训练进行学习. 而文献[62]是一种更具通用性的无监督学习域间映射的方法. 如图5(b)所示,CycleGAN定义两个映射$\boldsymbol{F}_{X\to Y}$和$\boldsymbol{F}_{Y\to X}$及相应的判别函数$\boldsymbol{D}_X$和$\boldsymbol{D}_Y$,通过对抗性损失使得生成的图像分布不断逼近目标域中数据分布. 同时,为防止生成器学习到具有欺骗性的造假数据,使用循环一致性(Cycle-consistency)损失. 值得注意的是,这类方法在应用于风格迁移任务时,除了能够迁移色彩、纹理等低级特征外,还能捕捉到笔触、语义等中高级特征. 尽管在风格化效果上取得了良好的表现,但在保留内容结构方面仍可能存在显著损失.

现有图像到图像翻译方法的另一个局限性是生成结果缺乏多样性,域和域之间的映射本质上应该是多模态的(multimodal). 为了处理多模态转换,一种直接的方法是将每种模态视作一个单独的域[63,64]. 另一类方法,比如MUNIT[65]和DRIT[66],应用对抗性损失将图像分解为域共享和域特定(domain-specific)的表示,如图5(c)所示. 在域特定的空间中学习到连续的分布,实现更丰富的生成结果.

4.2.3　基于StyleGAN实现风格迁移

受风格迁移领域启发,Karras等人[35]提出一种基于样式的生成框架StyleGAN. StyleGAN使用映射网络将

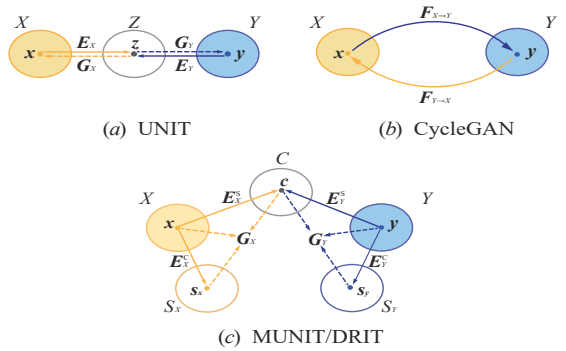


图5 无监督图像到图像翻译方法结构演进

从隐编码转换为隐空间中的向量送入生成器中，通过AdaIN变换控制生成结果，同时通过噪声增加生成多样性. StyleGAN的隐空间具备良好的可解释性，特别是对于具有较为固定结构特征的人脸图像，可以通过混合不同尺度的风格向量对图像语义特性进行控制，这对于肖像风格化等图像迁移任务是有利的.

然而，在数据不足的情况下，StyleGAN的泛化性能较弱，因此很多工作针对少样本学习展开研究. 举例来说，Ojha等人[67]使用跨域一致性正则项微调预训练模型实现域适应. 零样本或单样本学习是少样本学习的一个极端情况，即模型在没有或仅有一个训练样本的情况下，能够有效地学习并泛化到新任务上. 一些工作[68,69]通过将艺术风格图像反转到真实人脸域来构建跨域数据，实现单样本域适应.

此外，部分工作针对存在遮挡和配件的复杂风格迁移场景提出更通用的解决方案. 举例来说，Zhang等人[70]引入辅助网络增强实体生成. 而文献[71]通过校准目标域分布增强全局结构多样性，同时引入基于面部特征点的感知损失监督，鼓励具有夸张变形结构的风格化. 此类方法往往不要求精准的内容细节保留，而是以人脸身份一致性作为衡量指标.

4.3 基于扩散模型的风格迁移

本节回顾了基于扩散模型实现图像风格迁移的研究进展. 如图4所示，从传统扩散模型出发，深入到通过条件约束引导风格迁移的过程，最后聚焦于稳定扩散模型这一前沿变体为风格迁移领域带来的技术革新.

4.3.1 基于传统扩散模型实现风格迁移

去噪概率扩散模型(Denoising Diffusion Probabilistic Models，DDPM)[23]是从非平衡热力学中受到启发产生的隐变量模型，整体分为前向扩散过程和逆向去噪过程. 前向过程表示如式(6)所示. 从真实图像分布中随机采样一张图像$\boldsymbol{x}\sim q(\boldsymbol{x}_0)$，在$T$个步长逐步添加噪声，转换为预定义的高斯分布$\mathcal{N}(0,\boldsymbol{I})$.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}):=\mathcal{N}\left(\boldsymbol{x}_t;\ \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1},\beta_t\boldsymbol{I}\right) \tag{6}$$

由于状态转移分布$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$无法直接求得，DDPM使用参数化模型$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$模拟逆向扩散的推断过程，具体过程如式(7)所示. 其中，$\boldsymbol{\epsilon}_\theta$为U-Net架构的噪声预测网络，接受步长$t$和对应状态分布$\boldsymbol{x}_t$为输入，输出为所预测的加在$\boldsymbol{x}_t$上的噪声.

$$\begin{aligned}&p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)=\mathcal{N}\left(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t),\sigma_t\boldsymbol{I}\right),\\&\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)=\frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t-\frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t)\right)\end{aligned} \tag{7}$$

DDPM展现了令人印象深刻的图像生成能力，但生成一张图像需要经过多步迭代去噪. 为此，文献[72]将前向过程定义为非马尔可夫链，以牺牲较少图像质量为代价减少采样步数. 当随机噪声项参数σ_t为0时，反向去噪过程变成确定性映射，称为去噪扩散隐式模型(Denoising Diffusion Implicit Models，DDIM).

在此基础上，部分研究引入风格迁移领域特定的弱监督，对预训练的噪声预测模型$\boldsymbol{\epsilon}_\theta$进行整体微调，使得模型能够有效学习图像风格信息的修改能力，从而适应于风格迁移任务. 举例来说，文献[73]利用DDIM的确定性反演保留输入图像的内容结构，在CLIP方向损失监督下微调$\boldsymbol{\epsilon}_\theta$，对齐目标文本描述和生成的风格. 而Wang等人[74]显式去风格化地提取图像内容信息，并隐式地学到互补的风格信息. 利用CLIP空间的风格解纠缠损失微调$\boldsymbol{\epsilon}_\theta$，实现风格迁移. 然而此类方法需要针对每种新的风格参考对整个噪声预测模型进行微调，非常低效和耗时.

4.3.2 基于条件扩散模型实现风格迁移

为了增强图像生成的控制，条件扩散模型引入了分类器引导(Classifier-Guidance)和无分类器引导(Classifier-Free Guidance，CFG)两种策略.

分类器引导由文献[75]提出. 如式(8)所示，通过图片分类器的梯度调节反向扩散过程，在尽量保持图片生成多样性的前提下提升生成的准确性.

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,y)=\mathcal{N}\left(\boldsymbol{\mu}+s\boldsymbol{\Sigma}\nabla_{\boldsymbol{x}_t}\log p_\phi(y|\boldsymbol{x}_t),\boldsymbol{\Sigma}\right) \tag{8}$$

其中，分类器$p_\phi(y|\boldsymbol{x}_t)$能够灵活替换为其他引导机制(如不同判别器或损失函数)，优化扩散模型在去噪过程中的表现. 这一进展带来了图像风格迁移领域的突破，一些研究[18,76,77]利用从输入条件中提取的嵌入向量或网络特征来构建损失函数，以精确控制风格的融合过程. 文献[18]构建CLIP空间的方向损失函数引导隐空间变量的优化过程. 文献[76]使用生成样本和原始图像之间的逐patch对比损失控制图像生成的方向. 这种方法增加了对扩散模型的控制，但显式的分类器引导可能会带来局部空间集中问题，影响生成结果的多样性.

无分类器引导方法避免了显式分类器的限制，将

标签y送入模型参与前向训练,训练过程中并行优化基于条件的扩散模型与无条件的扩散模型,以生成既符合特定条件又保持多样性的样本.受此启发,文献[15,77,78]中皆使用CFG结合损失函数引导的方式实现风格迁移.

4.3.3　基于稳定扩散模型实现风格迁移

尽管条件控制的问题得到了解决,但DDPM和DDIM在像素空间进行操作,训练和推理代价较高.针对这一问题,稳定扩散模型(Stable Diffusion Models,SDM)[79]在隐空间而非像素空间中运行扩散过程,使推断速度更快.此外,SDM使用交叉注意力机制引入条件控制:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\cdot\boldsymbol{V} \tag{9}$$

$$\boldsymbol{Q}=\boldsymbol{W}_{\boldsymbol{Q}}\boldsymbol{\varphi}(z_t),\boldsymbol{K}=\boldsymbol{W}_{\boldsymbol{K}}\boldsymbol{c},\boldsymbol{V}=\boldsymbol{W}_{\boldsymbol{V}}\boldsymbol{c}$$

其中,$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$是学习到的线性投影变换,$\boldsymbol{\varphi}(z_t)$为采样深层特征,$\boldsymbol{c}$为嵌入条件.交叉注意力条件机制的引入使得稳定扩散模型具备多模态条件生成的能力,许多基于SDM的风格迁移算法研究陆续被提出.

首先,一些方法利用在大规模数据集上学到的通用特征,通过在特定任务数据集上进行少量训练,使模型适应新的特定任务或领域.举例来说,如图6(*a*)所示,Everaert等人[80]对稳定扩散模型的U-Net进行微调,通过修改初始隐编码分布实现对应目标风格化.然而,这类方法存在计算开销大且数据集收集困难的弊端.

如图6(*b*)所示,通过注意力修正操作来实现风格融合是更常见的手段.文献[81]首次揭示了文本到图像扩散模型中的交叉注意层所具有的语义控制能力,并通过修改交叉注意力层的交互过程实现了文本驱动的图像编辑,是注意力修正方面的开创性研究.基于此,文献[82,83]在去噪过程中保持原始输入图像的注意力图以保留内容结构,并通过注意力层交互实现风格转换.文献[16,84]结合自定义的适配器模块从参考风格提示中提取特征,通过交叉注意层将这些特征集成到扩散生成过程中,引导风格迁移过程.此外,Jeong等人[85]利用U-Net瓶颈层的语义隐空间进行渐进式内容注入,Chung等人[86]操纵SDM模型自注意力层特征进行风格融合,这些研究实现了无需训练的风格迁移方法.

另一类方法[87~89]关注提示工程(prompt engineering),将参考风格提示反演为文本嵌入,如图6(*c*)所示,充分利用SDM文本到图像生成能力实现风格化生成.其中,文献[88]提出了一种基于反演的方法InST,模型使用稳定扩散模型为网络主干,集成CLIP与多层交叉注意力机制,提取单个绘画图像的风格信息作为文本条件引导风格迁移.

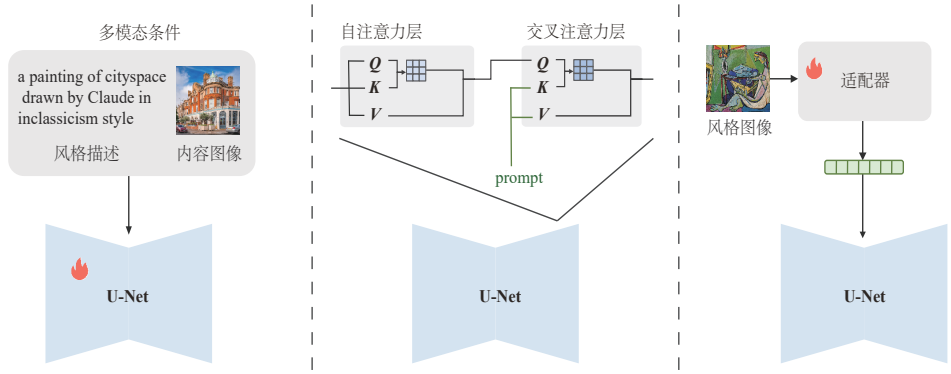


图6　基于稳定扩散模型实现风格迁移方法框架

4.4　基于其他模型架构的风格迁移

除上述3种主流架构,部分风格迁移方法基于其他模型,或是依据其特点设计了独特的网络架构.例如,早期的风格迁移方法[10,24]通过VGG网络的迭代优化实现风格转换.一些研究[90~92]应用基于流(flow)的模型.其中,文献[90]使用可逆流网络结合无偏风格转换模块进行风格传输,然而级联模块引入的冗余信息造成了伪影.文献[92]添加可逆残差网络模块改进可逆网络,并采用通道细化模块来避免冗余信息积累,解决了伪影问题.此外,Xia等人[93]设计了一种基于双边空间(bilateral space)的风格迁移算法,一路分支在低分辨率学习局部放射变换,另一路分支在原始分辨率进行风格映射.

4.5　小结

在现有的风格迁移方法中,不同模型架构之间存在显著的差异化亮点.首先,早期基于VGG的方法通过优化内容与风格之间的损失函数完成风格转移,面临计算开销大、生成速度慢的问题.基于自编码器架构的风格迁移方法实现了前馈式的任意风格迁移,并在引入注意力机制后显著提升了风格化效果,同时较好地保留了原内容图像的结构细节.另一方面,GAN和扩

散模型等生成模型的引入，进一步提升了风格迁移的多样性，展现出更强的表达能力，但在内容结构的保留上则有所牺牲. 其中，StyleGAN架构因其在人脸可解释性方面的优势，在肖像风格化领域展现了显著的应用潜力. 而稳定扩散模型展现出的文本到图像的生成能力推动了基于文本引导的风格迁移方法的发展，为风格迁移的个性化和多样化提供了更为丰富的控制手段. 未来的研究可以围绕跨模态引导展开，对如何提升风格迁移的灵活性和精准度进行深入探索.

5 实验比较与分析

5.1 数据集

图像风格迁移任务中所使用数据集分为两大类别：一类是内容图像数据集，提供基础图像内容；另一类是风格图像数据集，定义目标图像所需呈现的艺术风格. 根据不同任务场景，选择合适的数据集进行模型训练或测试，确保模型能准确执行特定风格迁移任务.

最常用的数据集包括MSCOCO[94]（118 287张训练图像，5 000张验证图像，涵盖80个对象类别），ImageNet[95]（1 400万张图像，涵盖生活中常见类别），以及LSUN[96]（900多万张图像，10个场景类别）. Places365[97]是一个场景识别数据集，涵盖434个场景类别. 对于肖像风格化任务，常采用FFHQ[98]和CelebA-HQ[99]等高质量人脸数据集，这些数据集在人脸属性及场景上具有丰富广泛的变化，有利于提升模型的泛化能力. 内容图像数据集的使用情况如表1所示[101~105].

表1 内容图像数据集汇总

名称	可用图片数量/幅	图像类型
WikiArt[101]	42 129	艺术风格迁移
PBN	81 752	艺术风格迁移
MetFaces[102]	1 336	肖像艺术风格化
AAHQ[103]	25 000	肖像艺术风格化
Danbooru Portraits	302 652	肖像动漫化
selfie2anime[104]	7 000	肖像动漫化
WebCaricature[105]	12 016	肖像漫画化

风格参考图像数据集如表2所示. 主要艺术风格参考来源包括WikiArt[106]. WikiArt网站是一个在线艺术作品数据库和虚拟美术馆，收录了来自各个时期、各个流派的艺术作品. 文献[101]基于该网站创建了一个包含42 129幅训练图像和10 628幅测试图像的精简数据集. PBN（Paint By Numbers）是Kaggle竞赛数据集，其中多数图像来自WikiArt. 对于照片级真实风格迁移任务，风格参考图像通常来自现实场景.

肖像风格化任务的风格参考涵盖艺术风格、漫画、动漫等多种类型. MetFaces[102]包含1 336幅艺术作品中的人脸图像（分辨率为1 024 × 1 024），AAHQ[103]从Artstation肖像频道收集了25 000张艺术脸部图像，Danbooru Portraits包含302 652张高质量动漫人脸图像，WebCaricature[105]数据集包含从网络收集的252个人的6 042幅漫画和5 974张照片，每张图片都标注了17个面部特征点.

表2 参考风格数据集汇总

名称	可用图片数量/幅	图像类型
MSCOCO[94]	164 037	场景和物体
ImageNet[95]	14 000 000+	场景和物体
LSUN[96]	9 000 000+	场景和物体
Places365[97]	10 000 000	场景
FFHQ[98]	70 000	肖像
CelebA[99]	200 000	肖像
CelebA-HQ[100]	30 000	肖像

5.2 评价体系

图像风格迁移任务的评价体系始终是一个复杂而关键的问题，当前这一领域的评估方法普遍从定性与定量两个维度来综合评估模型的性能. 定性评价往往依赖于观察者的主观审美判断，评价结果往往与个人文化背景、艺术偏好紧密相关，存在一定的主观性. 定量评估则通过客观指标来衡量模型性能.

表3列出了风格迁移任务中常用评价指标[107~111]的信息，覆盖了从内容相似度、风格化效果到用于视频风格迁移的时间一致性等多个方面. 同时对各个指标在本文提及的风格迁移研究中的使用率进行统计，下面对认可度较高的评价指标进行介绍.

5.2.1 内容保留

衡量内容保留程度的指标主要为结构相似性指数（Structural Similarity Index，SSIM）和学习感知图像块相似度（Learned Perceptual Image Patch Similarity，LPIPS）.

SSIM考虑了亮度、对比度和结构3个方面的差异，计算过程如式（10）所示：

$$\begin{aligned}\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y}) &= \left[l(\boldsymbol{x},\boldsymbol{y})^{\alpha}\cdot c(\boldsymbol{x},\boldsymbol{y})^{\beta}\cdot s(\boldsymbol{x},\boldsymbol{y})^{\gamma}\right],\\ l(\boldsymbol{x},\boldsymbol{y}) &= \frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1},\\ c(\boldsymbol{x},\boldsymbol{y}) &= \frac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2},\\ s(\boldsymbol{x},\boldsymbol{y}) &= \frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3}\end{aligned} \tag{10}$$

其中，$\boldsymbol{x}$和$\boldsymbol{y}$为两幅图像，μ、σ为均值和方差，σ_{xy}表示$\boldsymbol{x}$和$\boldsymbol{y}$之间的协方差. SSIM值范围为$[-1,1]$，越接近1表示图像越相似.

LPIPS使用预训练深度学习模型提取高级特征，并通过L_2距离比较特征差异. 相比像素级比较，LPIPS更接近于人类的视觉感知，值越小表示图像越相似.

表3　模型评价指标

指标名称	评估对象	计算方式	任务场景	使用率/%
SSIM[107] ↓	内容相似度	计算内容图和风格化结果在亮度、对比度和结构3个方面的差异	任意风格迁移	18.18
LPIPS[108] ↓	内容相似度	通过深度学习模型计算图像之间的感知差异	任意风格迁移	41.82
SC ↑	内容相似度	提取图像的语义分割掩码并计算差异	任意风格迁移	1.82
$\mathcal{L}_s$ ↓	风格	VGG中提取的风格特征差异	图像引导的风格迁移	27.27
FID[109] ↓	风格	图像集在Inception网络中的平均概率距离	图像引导的风格迁移	23.64
SIFID[110] ↓	风格	单幅图像FID	图像引导的风格迁移	5.45
ArtFID[111] ↓	风格	ArtFID=(1+LPIPS)(1+FID)	图像引导的风格迁移	3.64
风格欺骗率 ↓	风格	使用预训练风格分类模型进行Top-1分类准确率测试	图像引导的风格迁移	5.45
CLIP得分 ↑	风格	计算CLIP嵌入空间内文本提示与相应风格化图像之间的余弦相似度	文本引导的风格迁移	18.18
CLIP方向一致性 ↑	风格	分别计算源域和目标域间的文本和图像CLIP方向向量,计算余弦相似性	文本引导的风格迁移	5.45
时间/s ↑	速度	风格化所需平均时长	任意风格迁移	50.91
Diversity ↑	生成多样性	(1)对生成样本进行聚类并计算聚类内的平均LPIPS距离.(2)随机生成两个批次并计算批次之间的LPIPS距离	基于生成模型的风格迁移	1.82
身份相似性ID ↑	身份一致性	(1)利用预训练的人脸识别模型提取风格化前后的身份特征向量,通过余弦距离衡量相似度.(2)利用预训练模型提取人脸特征并计算余弦相似度.(3)利用预训练的分类网络对风格化图像进行Top-1分类准确率测试	肖像风格化	13.85
Temporal loss ↓	帧间一致性	利用光流信息计算前一帧与当前帧预测结果的一致性	视频风格迁移	9.09

5.2.2　风格化效果

衡量结果风格化效果的指标通常包括风格损失、FID(Fréchet Inception Distance)、CLIP分数3种.

风格损失通常基于风格化结果和风格参考图的格拉姆矩阵差异度量,计算过程如式(1)所示.另一种常用的风格损失通过衡量特征均值和方差的差异计算,如式(11)所示.其中,ϕ表示VGG中用于计算风格损失的层,通过比较风格化结果和风格特征的每通道均值μ与方差σ之间的差异衡量图像间风格差异.

$$\begin{aligned}\mathcal{L}_s=&\sum_{i=1}^{L}\left\|\mu\left(\phi_i(g)\right)-\mu\left(\phi_i(s)\right)\right\|_2\\&+\sum_{i=1}^{L}\left\|\sigma\left(\phi_i(g)\right)-\sigma\left(\phi_i(s)\right)\right\|_2\end{aligned}\tag{11}$$

FID评估不同风格化图像与目标风格图像之间的相似度,如式(12)所示.其中,μ和Σ分别是图像特征的均值与协方差矩阵.文献[110]提出针对单幅图像的评价方法SIFID.

$$\mathrm{FID}=\|\mu_r-\mu_g\|^2+\mathrm{Tr}\left(\Sigma_r+\Sigma_g-2(\Sigma_r\Sigma_g)^{\frac{1}{2}}\right)\tag{12}$$

CLIP分数用于评估基于多模态条件的风格迁移,通过比较生成图像与参考图像或文本的相似性计算得分.

此外,在风格迁移领域中广泛采用的一种量化评价方法是用户调研,即通过问卷等形式调查用户进行更客观的比较.尽管不同观察者可能对同一幅风格迁移后的图像评价迥异,用户调研方式通过大规模用户的参与尽可能消弭主观因素带来的影响,同时捕捉到算法难以量化的艺术审美层面,为风格迁移效果提供一个相对全面且贴近真实用户感受的评估视角.

5.3　评估与分析

本节选取部分具有代表性的风格迁移算法,在单个GeForce RTX 4090上进行实验,对模型性能进行进一步比较与分析.

5.3.1　图片引导的风格迁移算法分析

首先基于不同模型架构,选取了部分目前最先进的风格迁移算法,对其进行定性和定量的实验比较.实验中总共选用11张风格图像和30张内容图像,生成330个风格迁移结果.风格参考图像涵盖素描、油画、水彩等不同绘画类型,囊括印象主义、写实主义、立体主义等多种艺术流派.

图7展示了所选取的图像引导风格迁移算法的风格化结果.通过分析可以发现:

(1)基于AE方法容易造成伪影(例如图7第2~3行第4列).此外,AdaAttN[53]无法很好地捕获某些笔画模式,在某些平滑的区域产生重复扭曲的线条(例如图7第2行的第2、3、6列).AesPA-Net[54]通过考虑图案重复性获得更平滑的风格化结果,然而这种方法造成了内容结构上的损失.

(2)ArtFlow[90]使用流结构进行无偏风格传递,然而由于级联模块引入了冗余信息,在图像边缘产生了明显的伪影(如图7第4行的第3、5、6列).

(3)基于GAN的IEST[34]和CAST[112]都利用大规模艺术集的外部风格相关信息,更好地捕捉了风格参考的笔触信息.但其中IEST[34]使用二阶统计量作为风格表示,有时会出现颜色失真的情况(如图7第5行第2列).

(4)基于扩散模型的风格迁移算法相对来说更为耗时,风格化结果也更具有创造性. InST[88]和 DiffuseIT[18]方法都无法很好地保持图像内容结构(如图7第7~8行的第2、4、5列). 而StyleID[86]通过反向扩散过程保留原始内容的空间结构,表现出更优的性能.

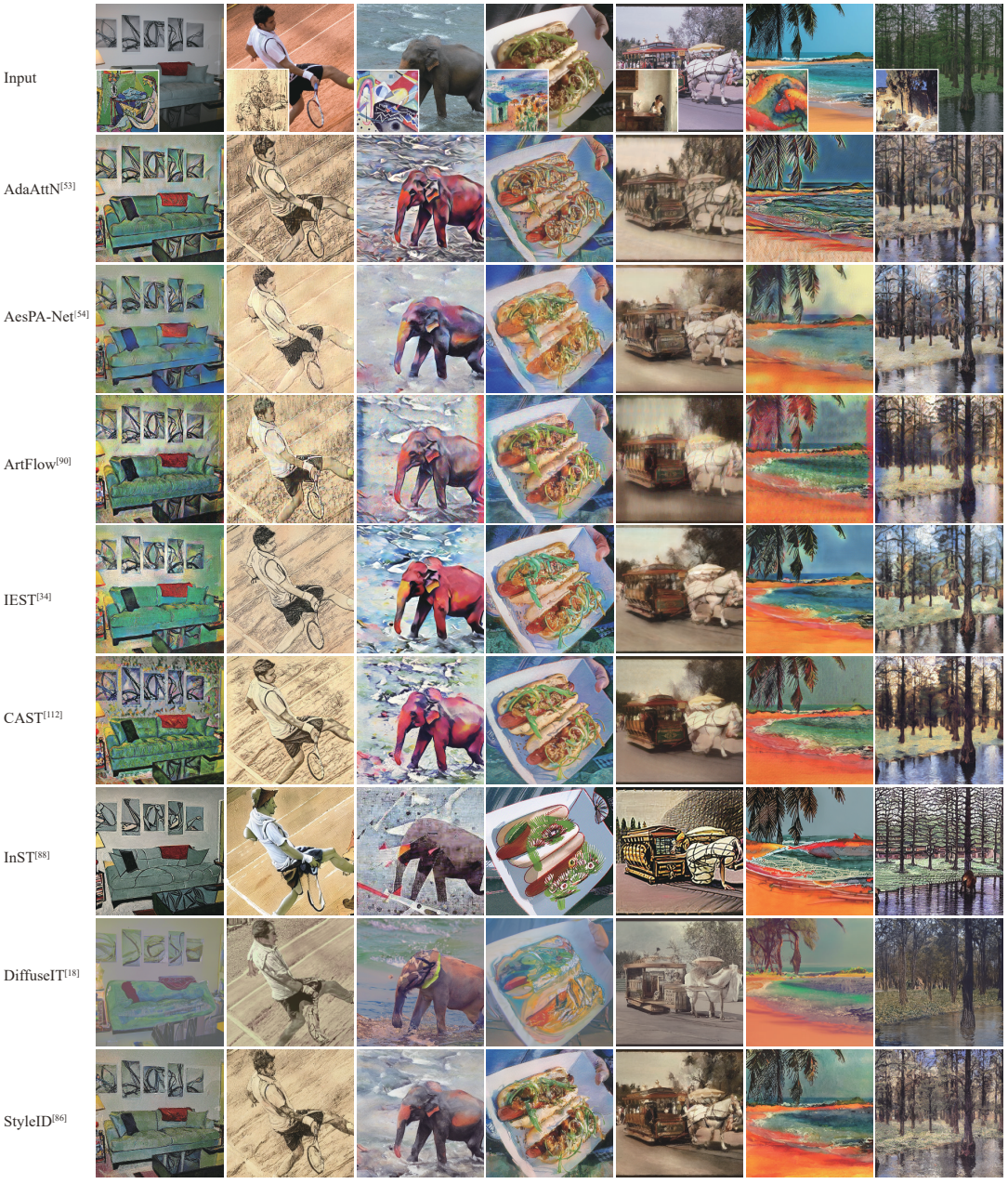


图7 图像引导的风格迁移算法的定性评估结果

表4中展示了部分定量指标结果. 从表4中可以看出,由于风格迁移没有统一客观的评价指标,这些方法在各类指标上的表现是参差不齐的. 举例来说,基于扩散模型的方法不依赖于风格损失 $\mathcal{L}_s$ 进行训练,因此 $\mathcal{L}_s$ 得分也相对更差,但InST[88]和DiffuseIT[18]分别取得了最优SIFID和次优FID得分. 总体来说,AdaAttN[53]和StyleID[86]在内容结构保持和风格化效果方面优于其他方法,而InST[88]和DiffuseIT[18]损失了较多的内容结构.

此外,为了探究模型的推理效率,对不同尺寸的图像进行风格迁移并记录各算法的平均推理时间,如表5所示. 相对来说,扩散模型通过去噪迭代进行采样,会消耗更多的时间. 而在扩散模型中,尽管

InST[88]的平均推理时间最少,但需要针对每一种新的风格参考进行新一轮时长约20 min的训练. DiffuseIT[18]则需要根据实际输入对隐变量进行迭代优化,耗时较长. 而StyleID[86]是一种无需训练的仅对注意力层进行修改的风格迁移方法,具有更高的实用价值.

表4　基于图像的风格迁移算法的定量评估结果

方法	SSIM ↑	LPIPS ↓	$\mathcal{L}_s$ ↓	FID ↓	SIFID ↓	ArtFID ↓	发表时间	模型架构
AdaAttN[53]	0.510 3	0.590 4	0.866 1	20.758 9	9.480 1	34.606 0	ICCV 2021	自编码器
AesPA-Net[54]	0.432 8	0.596 7	1.585 7	22.626 1	9.517 4	37.724 9	ICCV 2023	自编码器
ArtFlow[90]	0.395 1	0.524 8	1.389 4	23.758 7	9.486 9	37.752 7	CVPR 2021	流模型
IEST[34]	0.426 8	0.556 0	1.057 8	22.506 1	9.481 1	36.575 9	NeurIPS 2021	生成对抗网络
CAST[112]	0.479 1	0.577 7	1.562 9	22.693 6	9.479 2	37.383 1	SIGGRAPH 2022	生成对抗网络
InST[88]	0.352 5	0.586 6	5.489 4	24.361 6	9.439 8	40.239 6	CVPR 2023	扩散模型
DiffuseIT[18]	0.390 0	0.628 8	2.435 2	28.196 8	9.520 2	47.557 9	ICLR 2023	扩散模型
StyleID[86]	0.457 7	0.523 2	2.048 8	21.936 9	9.487 3	34.937 7	CVPR 2024	扩散模型

表5　基于图像的风格迁移算法的推理速度对比

方法	引导条件		推理时间/s		风格数
	内容	风格	256 × 256	512 × 512	
AdaAttN[53]	图像	图像	0.074 3	0.177 6	∞
AesPA-Net[54]	图像	图像	0.187 4	0.275 7	∞
ArtFlow[90]	图像	图像	0.045 0	0.085 0	∞
IEST[34]	图像	图像	0.005 3	0.005 1	∞
CAST[112]	图像	图像	0.004 5	0.005 7	∞
InST[88]	图像	图像	1.181 3	2.041 9	1
DiffuseIT[18]	图像	图像	32.256 4	48.900 6	1
StyleID[86]	图像	图像	3.245 9	28.413 4	∞
LDAST[36]	图像	文本	0.119 5	0.119 8	∞
ClipStyler[14]	图像	文本	15.350 5	30.485 5	1
DiffusionCLIP[73]	图像	文本	3.097 8	—	1
DEADiff[19]	文本/图像	文本/图像	—	1.655 4	∞
Textual Inversion[87]	文本	图像	0.162 3	0.413 9	1
InstantStyle[17](SDXL)	文本	图像	—	4.348 3	∞
InstantStyle[17](SD1.5)	文本	图像	0.692 2	0.977 8	∞

5.3.2　文本引导的风格迁移

如3.2节中所介绍的,文本引导的图像风格迁移算法可以分为两类:(1)以内容图像和风格文本描述为参考的风格迁移;(2)以文本提示和风格图像为参考的风格迁移. 因此,设置了两组实验,分别选取两类方法中具有代表性的模型,每组实验分别使用11个参考文本和11张参考图像获得121组风格迁移样本.

首先对以内容图像和风格文本描述为参考的图像风格迁移方法进行实验分析,图8展示了部分风格化结果,表6和表5中分别展示了部分定量指标和模型对不同尺寸图像的平均推理时间. 通过分析可以发现LDAST[36]具有最快的推理速度,同时在保留内容结构方面具有明显优势. 然而,如表6所示,LDAST[36]具有最低的CLIP得分. 一方面,这可能与该模型不是基于CLIP学习文本和图像对齐有关,但另一方面,该方法似乎只对颜色、纹理等特征进行迁移,在绘画轮廓上没有呈现出我们所期望的风格(如图8第3行第2、4列). ClipStyler[14]和DiffusionCLIP[73]在CLIP方向损失的监督下进行训练,可能会在CLIP误导下生成不相关的图案和伪影(如图8第4~5行第2列). 其中,DiffusionCLIP[73]的生成结果更加自然,这也表明扩散模型比自编码器能获得更加逼真的结果. 此外,从表5中可以看出,这两种方法都需要较长的推理时间. 而DEADiff[19]通过提示工程完成风格注入,尽管获得了更自然且富有艺术感的风格化结果,但无法保留内容结构,甚至会完全丢失图像内容(如图8第6行第4列). 相应地,DEADiff[19]也获得了最差的SSIM和LPIPS分数.

图9展示了以内容文本提示和风格图像为参考的风格迁移算法的风格化结果. 对其进行分析发现,Textual Inversion[87]在部分风格化结果中会出现风格参考中的语义对象(如图9第3行的第1、5列),而

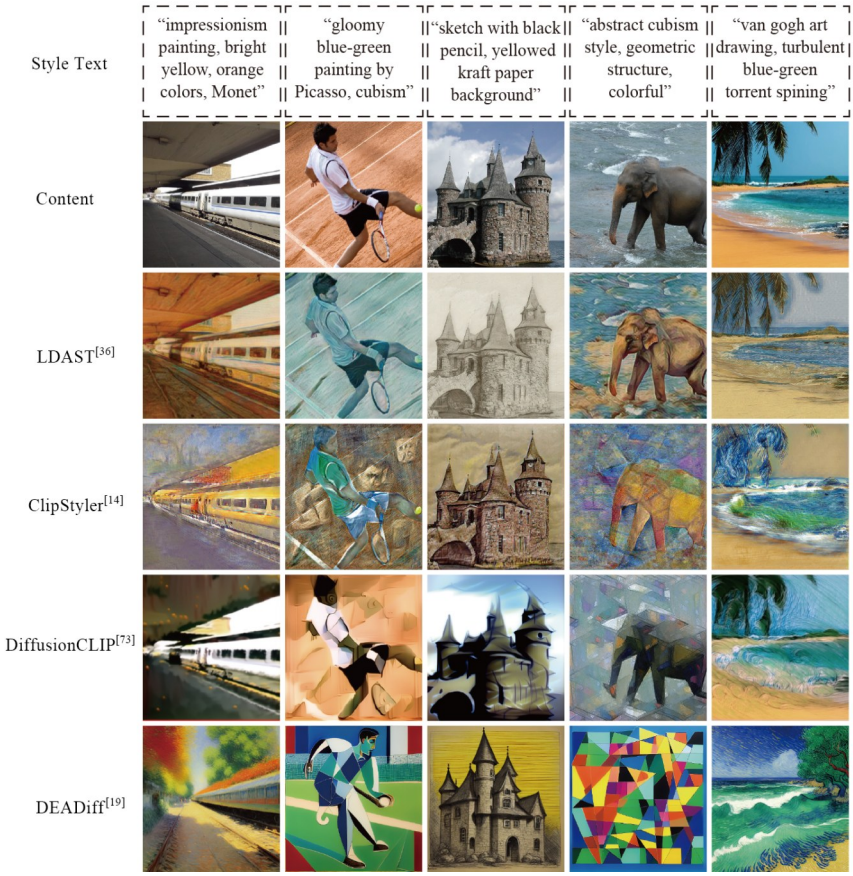


图 8 以内容图像和风格文本描述为参考的图像风格迁移结果

表 6 以内容图像和风格文本描述为参考的风格迁移方法定量评估结果

方法	SSIM ↑	LPIPS ↓	S_{Clip} ↑	发表时间	模型架构
LDAST[36]	0.476 2	0.573 7	0.228 0	ECCV 2022	生成对抗网络
ClipStyler[14]	0.297 1	0.459 2	0.341 8	CVPR 2022	自编码器
DiffusionCLIP[73]	0.532 4	0.614 3	0.268 6	CVPR 2022	扩散模型
DEADiff[19]	0.200 1	0.816 0	0.276 0	CVPR 2024	扩散模型

DEADiff[19]和 InstantStyle[17]的生成结果与风格参考图像在风格纹理、笔触上差异较大，并没有展现出期望的素描线稿及毕加索抽象画等绘画风格. 这反映出目前解耦风格参考图像中的风格和语义信息的手段尚未成熟，仍然有值得探索的空间. 表 7 显示了定量分析的结果，其中，CLIP 得分越高意味着输出结果和参考内容文本更加匹配，其余 3 个指标则衡量了图像的风格化效果. 总体看来，InstantStyle[17]具有最高的 CLIP 得分，实现了生成图像内容与输入文本的高度匹配，同时取得了较好的风格化效果. 从推理速度方面来看，如表 5 所示，InST[88]平均推理速度最快，但是每个预训练好的模型只能实现某种特定风格化. 综合各指标考量，可以发现 InstantStyle[17]在文本内容对齐、风格化效果和推理速度之间取得了较好的权衡.

进一步地，此类方法可以结合 ControlNet 进行辅助控制，实现保留内容结构的风格迁移. ControlNet 使得除了文本描述外，模型还可以接受分割图像、边缘检测结果或颜色标注等其他多模态形式的条件输入，从而使生成的图像更符合用户的特定需求. 类似地，InstantStyle[17]通过结合文本、图像特征等多种条件引导图像生成，使得风格迁移更加灵活和精准. 因此，在以风格参考图像与内容文本作为输入，执行文本至图像的特定风格化生成过程中，可以通过增加内容图像作为输入，提取其边缘信息，并利用 ControlNet 技术保留这些边缘结构，使得生成结果的同时具有内容图像的结构和风格图像的风格笔触. 图 9 右半部分展示了 InstantStyle[17]在使用 ControlNet 辅助控制下进行风格迁移的结果. 可以看出，图像保留了内容图像的边缘结构，同时具有和风格参考图像类似的色彩、笔触等绘画风格.

除此以外，还针对肖像风格化任务进行了实验对比与分析. 实验中选取 4 种先进的肖像风格化方法，使用 8 张风格图像和 8 张内容图像进行风格迁移，风格参考图囊括 3D 动画、夸张漫画式、卡通等多种类型. 图 10 展示了部分风格化结果，表 8 展示了部分定量指标. 通

图 9　以内容文本提示和风格图像为参考的图像风格迁移结果

表 7　以内容文本提示和风格图像为参考的风格迁移方法定量评估结果

方法	$\mathcal{L}_s\downarrow$	FID ↓	SIFID ↓	$S_{\text{Clip}}\uparrow$	发表时间	模型架构
Textual Inversion[87]	2.327 8	22.256 3	9.483 800	0.247 8	Arxiv 2022	扩散模型
DEADiff[19]	9.695 8	25.473 9	9.443 900	0.275 6	CVPR 2024	扩散模型
InstantStyle[17]	2.135 8	25.718 6	9.580 827	0.276 4	Arxiv 2024	扩散模型

过观察表 8 可以发现,DualStyleGAN[113]在风格一致性方面表现最优,同时在人脸身份一致性方面表现较差.这可能是由于DualStyleGAN[113]在大量同类型风格数据集上进行训练,能更准确地捕捉到风格特征,但另一方面,这可能会使得风格化结果更接近训练集的平均水平而非输入的参考风格图(如图 10 第 4 行第 5、6 列).JoJoGAN[69]和 ZePo[114]基本保留了原始内容结构,同时融入了风格图的线条感、色彩搭配或纹理信息,但在有夸张形变的风格化方面表现较差(如图 10 第 3、6 行第 3、4 列),而文献[115]的方法在具有夸张漫画形变的风格化方面具有更好的表现.

6　未来研究方向

尽管图像风格迁移领域已经取得了显著进展,但现有方法仍存在诸多挑战,未来的研究可以关注以下 5 个方向.

(1)语义与风格的解耦.在风格迁移任务中,如何准确解耦图像中的语义和风格信息一直是一个关键问题.尽管当前图像引导的风格迁移技术已展现出相对成熟的解决策略,但在跨模态的应用场景中,风格与语义信息的解耦难题依旧亟待突破.当前一些基于 CLIP 的文本引导方法虽能捕捉图像和文本的对齐关系,但其内在机制却难以彻底区分风格与语义层面,风格化结果中会出现风格图像中的语义对象.另一方面,若对风格特征的提取与表征能力不足,则难以充分传递原风格图像中蕴含的艺术特征.最近一些研究已经着手从 CLIP 空间将内容和风格进行解耦[116],或者针对生成模型提出新框架来理解和提取图像中的风格描述符[117].进一步优化跨模态信息中的风格与语义解耦技术,将是提升图像生成视觉质量的关键所在.

(2)统一评价标准.由于审美的主观性,图像风格迁移领域缺乏一个普遍认可且客观统一的评价标准.不同研究团队往往依据各自的研究视角与需求,采用不同甚至自定义的评估指标,如风格损失、感知损失等.这些指标虽各有侧重,但难以全面、客观地反映风格迁移效果的好坏.特别是在文本引导的风格迁移中,由于 CLIP 模型无法解耦风格和语义层面,该指标是不足以评估跨模态信息间的内容与风格相似度的.因此,如何建立一个既能准确反映风格迁移质量又能考虑用户主观感受的统一评价标准是值得关注的问题.

(3)视频风格迁移中的时序一致性.与图像风格迁移相比,视频风格迁移不仅需要保持每帧的风格化效果,还要确保生成视频的时序一致性和运动平滑性.当前的方法通过引入光流误差等正则项来保证时间一致性,然而这些方法无法逼真地传递艺术风格的纹理和笔触等信息,使得风格化质量下降.未来研究可以探索在不牺牲风格化质量的前提下,提升视频风格迁移的稳定性和一致性.

(4)3D 风格迁移.尽管风格迁移技术在二维图像风格化领域已取得显著进展,但这些方法在 3D 场景中

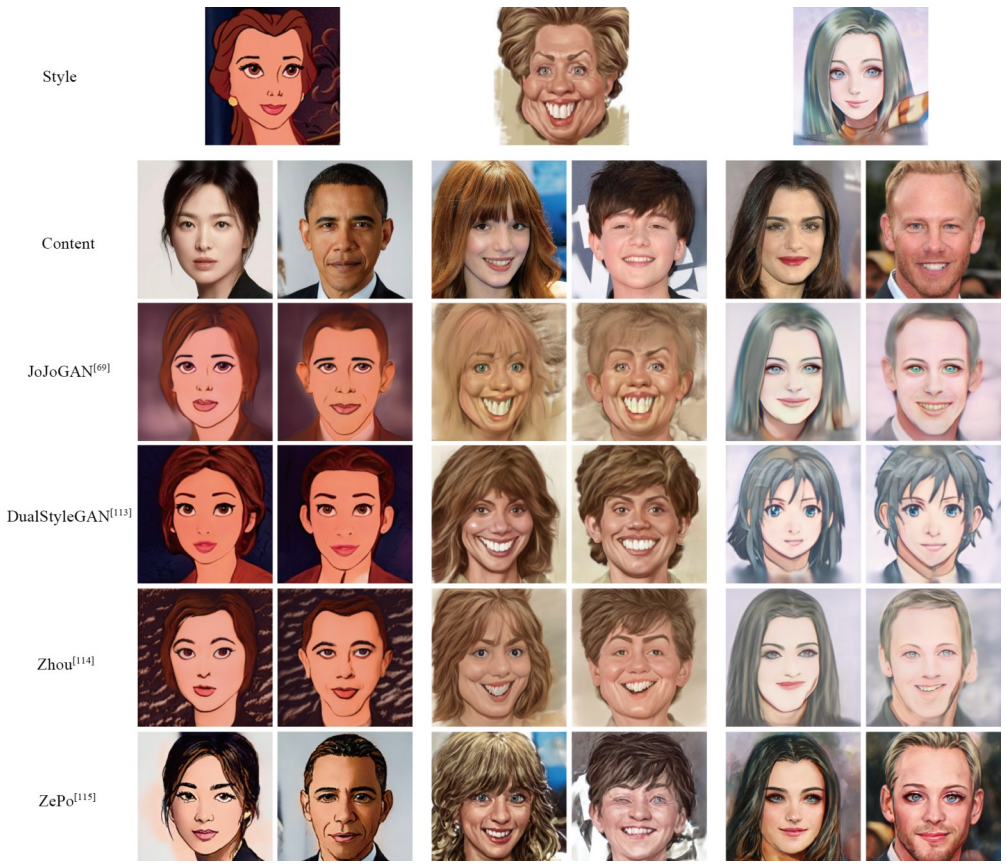


图 10 肖像风格化结果

表 8 肖像风格化方法定量评估结果

方法	$\mathcal{L}_s\downarrow$	SIFID↓	ID↑	发表时间	模型架构
JoJoGAN[69]	0.348 2	2.911 2	0.274 8	ECCV 2022	生成对抗网络
DualStyleGAN[113]	0.185 3	1.928 5	0.145 6	CVPR 2022	生成对抗网络
Zhou[114]	0.263 4	2.924 9	0.340 5	CVPR 2024	生成对抗网络
ZePo[115]	0.421 4	2.920 2	0.294 8	ACM MM24	扩散模型

的拓展仍处于未充分探索的阶段. 目前已有部分研究[118~122]开始涉足这一领域,尝试将风格迁移技术应用于3D对象或场景中,这不仅是在技术上的探索,也是一项富有创意的实验. 3D场景下的风格化旨在根据给定的风格图像,从任意视角生成风格化图像,并且在从不同视点进行渲染时确保风格化效果的一致性与连贯性. 这一任务不仅要求技术能够捕捉并应用复杂的风格特征,还需保证在三维空间中的视角变换不会破坏风格的统一性和场景的逼真度.

(5)多模态大语言模型驱动的风格迁移. 随着多模态大语言模型(Multimodal Large Language Models, MLLMs)的兴起,图像风格迁移迎来了新的机遇. MLLMs以传统大语言模型为骨干网络,在继承其强大的语义理解和推理能力的同时,具备接收、理解以及生成其他模态的能力. 此外,MLLMs的出现使得图像生成方式趋于更灵活的交互式生成,通过与用户进行对话指导迭代的图像生成或编辑,使得生成图像更令人满意. 然而,这也带来了新的挑战:首先,MLLMs在风格迁移任务中的泛化能力尚未验证,如何高效融合视觉与文本模态的风格表征仍待探索;其次,庞大计算成本可能限制实时应用,需研究轻量化或蒸馏技术以提升效率. 当前已有学者[123~126]开始探索基于MLLMs的个性化图像生成技术,相信随着MLLMs技术的发展,图像风格迁移领域也将实现新的突破.

7 结论

风格迁移算法将给定的艺术风格参考与另一幅图的内容结合,创造出全新的视觉效果,这一图像编辑技术在过去几年受到研究者的广泛关注. 本文主要总结了基于神经网络的图像风格迁移方法,并对这些方法分别从引导条件的角度和网络架构的角度进行分类和描述. 依据引导条件将风格迁移算法划分为图像引导的风格迁移方法和文本引导的风格迁移算法两类,介绍了基于自编码器、生成对抗网络、扩散模型和其他模型架构的风格迁移方法. 随后,对相关数据集和评价体系进行介绍,并对部分现有方法进行实验评估与分析. 最后,对风格迁移任务未来研究方向进行展望,相信未

来的图像风格迁移技术将能够突破现有局限,为艺术创作、视觉设计等领域带来更加丰富、多样的可能性.

参考文献

[1] 李宝奇, 黄海宁, 刘纪元, 等. 基于改进CycleGAN的光学图像迁移生成水下小目标合成孔径声纳图像算法研究[J]. 电子学报, 2021, 49(9): 1746-1753.
LI B Q, HUANG H N, LIU J Y, et al. Optical image-to-underwater small target synthetic aperture sonar image translation algorithm based on improved CycleGAN[J]. Acta Electronica Sinica, 2021, 49(9): 1746-1753. (in Chinese)

[2] 杨曦, 张鑫, 郭浩远, 等. 基于不变特征的多源遥感图像舰船目标检测算法[J]. 电子学报, 2022, 50(4): 887-899.
YANG X, ZHANG X, GUO H Y, et al. Invariant features based ship detection model for multi-source remote sensing images[J]. Acta Electronica Sinica, 2022, 50(4): 887-899. (in Chinese)

[3] LI Y M, ZHANG D, KEUPER M, et al. Intra-source style augmentation for improved domain generalization[C]// 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2023: 509-519.

[4] JIA Y R, HOYER L, HUANG S Y, et al. DGInStyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 91-109.

[5] LI H L, LI W, CAO H, et al. Unsupervised domain adaptation for face anti-spoofing[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(7): 1794-1809.

[6] 李大锦, 高文冉, 高俊杰. 基于kuwahara滤波的视频风格化框架[J]. 电子学报, 2020, 48(3): 538-544.
LI D J, GAO W R, GAO J J. Artistic video stylization based on kuwahara filter[J]. Acta Electronica Sinica, 2020, 48(3): 538-544. (in Chinese)

[7] HERTZMANN A, JACOBS C E, OLIVER N, et al. Image analogies[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. New York: ACM, 2023: 557-570.

[8] WANG N N, TAO D C, GAO X B, et al. Transductive face sketch-photo synthesis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(9): 1364-1376.

[9] GATYS L A, ECKER A S, BETHGE M, et al. Texture synthesis using convolutional neural networks[C]//Advances in Neural Information Processing Systems. Washington: American Chemical Society, 2015: 262-270.

[10] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2414-2423.

[11] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. New York: PMLR, 2021: 8748-8763.

[12] PATASHNIKO,WUZZ,SHECHTMANE,etal.StyleCLIP: Text-driven manipulation of StyleGAN imagery[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 2065-2074.

[13] GAL R, PATASHNIK O, MARON H, et al. StyleGAN-NADA[J]. ACM Transactions on Graphics, 2022, 41(4): 1-13.

[14] KWON G, YE J C. CLIPstyler: Image style transfer with a single text condition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18041-18050.

[15] HUANG N S, ZHANG Y X, TANG F, et al. DiffStyler: Controllable dual diffusion for text-driven image stylization[EB/OL]. (2023-12-18)[2025-04-08]. https://arxiv.org/abs/2211.10682v2.

[16] CHEN D Y, TENNENT H, HSU C W. ArtAdapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 8619-8628.

[17] WANG H F, SPINELLI M, WANG Q X, et al. InstantStyle: Free lunch towards style-preserving in text-to-image generation[EB/OL]. (2024-04-04) [2025-04-08]. https://arxiv.org/abs/2404.02733v2.

[18] PENG D, HU P, KE Q H, et al. Diffusion-based image translation with label guidance for domain adaptive semantic segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 808-820.

[19] QI T H, FANG S C, WU Y Z, et al. DEADiff: An efficient stylization diffusion model with disentangled representations[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 8693-8702.

[20] JING Y C, YANG Y Z, FENG Z L, et al. Neural style transfer: A review[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(11): 3365-3385.

[21] CAI Q, MA M X, WANG C, et al. Image neural style

transfer: A review[J]. Computers and Electrical Engineering, 2023, 108: 108723.

[22] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. (2014-06-10)[2025-04-08]. https://arxiv.org/abs/1406.2661v1.

[23] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in neural information processing systems. New York: NIPS, 2020: 6840-6851.

[24] LI C, WAND M. Combining Markov random fields and convolutional neural networks for image synthesis[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2479-2486.

[25] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 1510-1519.

[26] LI Y J, FANG C, YANG J M, et al. Universal style transfer via feature transforms[EB/OL]. (2017-11-17) [2025-04-08]. https://arxiv.org/abs/1705.08086v2.

[27] LUAN F J, PARIS S, SHECHTMAN E, et al. Deep photo style transfer[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6997-7005.

[28] LI Y J, LIU M Y, LI X T, et al. A closed-form solution to photorealistic image stylization[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 468-483.

[29] YOO J, UH Y, CHUN S, et al. Photorealistic style transfer via wavelet transforms[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 9035-9044.

[30] CAO K D, LIAO J, YUAN L. CariGANs: Unpaired photo-to-caricature translation[EB/OL]. (2018-11-02) [2025-04-08]. https://arxiv.org/abs/1811.00222v2.

[31] SHI Y C, DEB D, JAIN A K. WarpGAN: Automatic caricature generation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 10754-10763.

[32] SANAKOYEU A, KOTOVENKO D, LANG S, et al. A style-aware content loss for real-time HD style transfer[C]// Computer Vision-CCV 2018. Cham: Springer International Publishing, 2018: 715-731.

[33] KOTOVENKO D, SANAKOYEU A, MA P C, et al. A content transformation block for image style transfer[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 10032-10041.

[34] CHEN H B, ZHAO L, WANG Z Z, et al. Artistic style transfer with internal-external learning and contrastive learning[C]//Advances in Neural Information Processing Systems. San Diego: NIPS, 2021: 26561-26573.

[35] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4401-4410.

[36] FU T J, WANG X E, WANG W Y. Language-Driven artistic style transfer[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 717-734.

[37] ZHANG L M, RAO A Y, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 3813-3824.

[38] WANG Z X, WANG X T, XIE L B, et al. StyleAdapter: A unified stylized image generation model[EB/OL]. (2024-10-30)[2025-04-08]. https://arxiv.org/abs/2309.01770v2.

[39] CUI X, LI Z K, LI P P, et al. INSTASTYLE: Inversion noise of a stylized image is secretly a style adviser[C]// Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 455-472.

[40] LI W, FANG M Y, ZOU C, et al. StyleTokenizer: Defining image style by a single instance for controlling diffusion models[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 110-126.

[41] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[42] JOHNSON J, ALAHI A, LI F F. Perceptual losses for real-time style transfer and super-resolution[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 694-711.

[43] DUMOULIN V, SHLENS J, KUDLUR M. A learned representation for artistic style[EB/OL]. (2017-02-09)[2025-04-08]. https://arxiv.org/abs/1610.07629v5.

[44] CHEN D D, YUAN L, LIAO J, et al. StyleBank: An explicit representation for neural image style transfer[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2770-2779.

[45] GHIASI G, LEE H, KUDLUR M, et al. Exploring the structure of a real-time, arbitrary neural artistic stylization

network[EB/OL]. (2017-08-24) [2025-04-08]. https://arxiv.org/abs/1705.06830v2.

[46] LI X T, LIU S F, KAUTZ J, et al. Learning linear transformations for fast image and video style transfer[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3809-3817.

[47] CHEN T Q, SCHMIDT M. Fast patch-based style transfer of arbitrary style[EB/OL]. (2016-12-13) [2025-04-08]. https://arxiv.org/abs/1612.04337v1.

[48] SHENG L, LIN Z Y, SHAO J, et al. Avatar-net: Multi-scale zero-shot style transfer by feature decoration[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8242-8250.

[49] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 21-25.

[50] PARK D Y, LEE K H. Arbitrary style transfer with style-attentional networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 5873-5881.

[51] YAO Y, REN J Q, XIE X S, et al. Attention-aware multi-stroke style transfer[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1467-1475.

[52] DENG Y Y, TANG F, DONG W M, et al. StyTr2: Image style transfer with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11316-11326.

[53] LIU S H, LIN T W, HE D L, et al. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 6629-6638.

[54] HONG K, JEON S, LEE J, et al. AesPA-net: Aesthetic pattern-aware style transfer networks[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 22701-22710.

[55] LI C, WAND M. Precomputed real-time texture synthesis with markovian generative adversarial networks[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 702-716.

[56] MIRZA M. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2025-04-08]. https://arxiv.org/abs/1411.1784.

[57] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 5967-5976.

[58] LIU M Y, TUZEL O. Coupled generative adversarial networks[EB/OL]. (2016-09-20)[2025-04-08]. https://arxiv.org/abs/1606.07536v2.

[59] LIU M Y, BREUEL T, KAUTZ J. Unsupervised image-to-image translation networks[C]//Advances in neural information processing systems. New York: NIPS, 2017: 700-708.

[60] KIM T, CHA M, KIM H, et al. Learning to discover cross-domain relations with generative adversarial networks[C]//International Conference on Machine Learning. New York: PMLR, 2017: 1857-1865.

[61] YI Z L, ZHANG H, TAN P, et al. DualGAN: Unsupervised dual learning for image-to-image translation[C]// 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2868-2876.

[62] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2242-2251.

[63] CHOI Y, CHOI M, KIM M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8789-8797.

[64] ANOOSHEH A, AGUSTSSON E, TIMOFTE R, et al. ComboGAN: Unrestrained scalability for image domain translation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2018: 783-790.

[65] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 179-196.

[66] LEE H Y, TSENG H Y, MAO Q, et al. DRIT++: Diverse image-to-image translation via disentangled representations[J]. International Journal of Computer Vision, 2020, 128(10): 2402-2417.

[67] OJHA U, LI Y J, LU J W, et al. Few-shot image generation via cross-domain correspondence[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 10743-10752.

[68] ZHU P H, ABDAL R, FEMIANI J, et al. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks[EB/OL]. (2021-11-28) [2025-04-09]. https://arxiv.org/abs/2110.08398v2.

[69] CHONG M J, FORSYTH D. JoJoGAN: One shot face stylization[C]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 128-152.

[70] ZHANG Z, LIU Y, HAN C, et al. Generalized one-shot domain adaptation of generative adversarial networks[C]// Advances in Neural Information Processing Systems. Beijing: University of Chinese Academy of Sciences, 2022: 13718-13730.

[71] MEN Y F, YAO Y, CUI M M, et al. DCT-net[J]. ACM Transactions on Graphics, 2022, 41(4): 1-9.

[72] SONG J M, MENG C L, ERMON S. Denoising diffusion implicit models[EB/OL]. (2022-10-05) [2025-04-08]. https://arxiv.org/abs/2010.02502v4.

[73] KIM G, KWON T, YE J C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 2416-2425.

[74] WANG Z Z, ZHAO L, XING W. StyleDiffusion: Controllable disentangled style transfer *via* diffusion models[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 7643-7655.

[75] DHARIWAL P, NICHOL A. Diffusion models beat GANs on image synthesis[C]//Advances in Neural Information Processing Systems. San Diego: NIPS, 2021: 8780-8794.

[76] YANG S, HWANG H, YE J C. Zero-shot contrastive loss for text-guided diffusion image style transfer[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 22816-22825.

[77] HUANG N S, TANG F, DONG W M, et al. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1085-1094.

[78] CHO H, LEE J, CHANG S, et al. One-shot structure-aware stylized image synthesis[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 8302-8311.

[79] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10674-10685.

[80] EVERAERT M N, BOCCHIO M, ARPA S, et al. Diffusion in style[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 2251-2261.

[81] HERTZ A, MOKADY R, TENENBAUM J, et al. Prompt-to-prompt image editing with cross attention control[EB/OL]. (2022-08-02)[2025-04-08]. https://arxiv.org/abs/2208.01626v1.

[82] LI S M, VAN DE WEIJER J, HU T H, et al. StyleDiffusion: Prompt-embedding inversion for text-based editing[EB/OL]. (2024-12-06)[2025-04-08]. https://arxiv.org/abs/2303.15649v3.

[83] PARMAR G, KUMAR SINGH K, ZHANG R, et al. Zero-shot image-to-image translation[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings. New York: ACM, 2023: 1-11.

[84] WANG Q X, BAI X, WANG H F, et al. InstantID: Zero-shot identity-preserving generation in seconds[EB/OL]. (2024-02-02)[2025-04-08]. https://arxiv.org/abs/2401.07519v2.

[85] JEONG J, KWON M, UH Y. Training-free content injection using h-space in diffusion models[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2024: 5139-5149.

[86] CHUNG J, HYUN S, HEO J P. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 8795-8805.

[87] GAL R, ALALUF Y, ATZMON Y, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion[EB/OL]. (2022-08-02) [2025-04-08]. https://arxiv.org/abs/2208.01618v1.

[88] ZHANG Y X, HUANG N S, TANG F, et al. Inversion-based style transfer with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 10146-10156.

[89] KAWAR B, ZADA S, LANG O, et al. Imagic: Text-based real image editing with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 6007-6017.

[90] AN J, HUANG S Y, SONG Y B, et al. ArtFlow: Unbiased image style transfer via reversible neural flows[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021:

862-871.

[91] FAN W C, CHEN J H, MA J B, et al. StyleFlow for content-fixed image to image translation[EB/OL]. (2022-07-05)[2025-04-08]. https://arxiv.org/abs/2207.01909v1.

[92] WEN L F, GAO C Y, ZOU C Q. CAP-VSTNet: Content affinity preserved versatile style transfer[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 18300-18309.

[93] XIA X D, ZHANG M, XUE T F, et al. Joint bilateral learning for real-time universal photorealistic style transfer[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 327-342.

[94] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.

[95] 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. Piscataway: IEEE, 2009: 248-255.

[96] YU F, SEFF A, ZHANG Y D, et al. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop[EB/OL]. (2016-06-04) [2025-04-08]. https://arxiv.org/abs/1506.03365v3.

[97] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: A 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464.

[98] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4401-4410.

[99] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 3730-3738.

[100] KARRAS T, AILA T M, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[EB/OL]. (2018-02-26) [2025-04-08]. https://arxiv.org/abs/1710.10196v3.

[101] 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, 2019, 28(1): 394-409.

[102] KARRAS T, AITTALA M, HELLSTEN J, et al. Training generative adversarial networks with limited data[C]//Advances in Neural Information Processing Systems. New York: NIPS, 2020: 12104-12114.

[103] LIU M C, LI Q, QIN Z K, et al. BlendGAN: Implicitly GAN blending for arbitrary stylized face generation[EB/OL]. (2021-10-22) [2025-04-08]. https://arxiv.org/abs/2110.11728v1.

[104] KIM J, KIM M, KANG H, et al. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation[EB/OL]. (2020-04-08)[2025-04-08]. https://arxiv.org/abs/1907.10830v4.

[105] HUO J, LI W B, SHI Y H, et al. WebCaricature: A benchmark for caricature recognition[EB/OL]. (2018-08-09)[2025-04-08]. https://arxiv.org/abs/1703.03230v4.

[106] PHILLIPS F, MACKINTOSH B. Wiki art gallery, inc.: A case for critical thinking[J]. Issues in Accounting Education, 2011, 26(3): 593-608.

[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] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 586-595.

[109] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[EB/OL]. (2018-01-12)[2025-04-08]. https://arxiv.org/abs/1706.08500v6.

[110] SHAHAM T R, DEKEL T, MICHAELI T. SinGAN: Learning a generative model from a single natural image[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 4570-4580.

[111] WRIGHT M, OMMER B. ArtFID: Quantitative evaluation of neural style transfer[C]//Pattern Recognition. Cham: Springer International Publishing, 2022: 560-576.

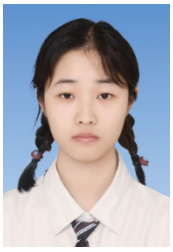
[112] ZHANG Y X, TANG F, DONG W M, et al. Domain enhanced arbitrary image style transfer via contrastive learning[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings. New York: ACM, 2022: 1-8.

[113] YANG S, JIANG L M, LIU Z W, et al. Pastiche master: Exemplar-based high-resolution portrait style transfer[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022:

7683-7692.
[114] LIU J, HUANG H B, CAO J, et al. ZePo: Zero-shot portrait stylization with faster sampling[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 3509-3518.
[115] ZHOU Y, CHEN Z C, HUANG H. Deformable one-shot face stylization via DINO semantic guidance[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 7787-7796.
[116] CAI Y C, LIU Y H, ZHANG Z, et al. CLAP: Isolating content from style through contrastive learning with augmented prompts[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 130-147.
[117] SOMEPALLI G, GUPTA A, GUPTA K, et al. Investigating Style similarity in diffusion models[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 143-160.
[118] KOTOVENKO D, GREBENKOVA O, SARAFIANOS N, et al. WaSt-3D: Wasserstein-2 distance for scene-to-scene stylization on 3D Gaussians[C]//Computer Vision-ECCV 2024. Cham: Springer Nature Switzerland, 2024: 298-314.
[119] LIU K H, ZHAN F N, XU M Y, et al. StyleGaussian: Instant 3D style transfer with Gaussian splatting[C]// SIGGRAPH Asia 2024 Technical Communications. New York: ACM, 2024: 1-4.
[120] CHEN Z, XU X D, YAN Y C, et al. HyperStyle3D: Text-guided 3D portrait stylization via hypernetworks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 9997-10010.
[121] SONG G X. AgileGAN3D: Few-shot 3D portrait stylization by augmented transfer learning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2024: 765-774.
[122] CHEN Y S, YUAN Q, LI Z Q, et al. UPST-NeRF: Universal photorealistic style transfer of neural radiance fields for 3D scene[J]. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(4): 2045-2057.
[123] DONG R P, HAN C R, PENG Y A, et al. DreamLLM: Synergistic multimodal comprehension and creation[EB/OL]. (2024-03-15) [2025-04-08]. https://arxiv. org/abs/2309.11499v2.
[124] ZHOU Y F, ZHANG R Y, GU J X, et al. Customization assistant for text-to-image generation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 9182-9191.
[125] ZHENG K Z, HE X H, WANG X E. MiniGPT-5: Interleaved vision-and-language generation *via* generative vokens[EB/OL]. (2024-03-15) [2025-04-08]. https://arxiv.org/abs/2310.02239v3.
[126] GE Y Y, ZHAO S J, ZENG Z Y, et al. Making LLaMA SEE and draw with SEED tokenizer[EB/OL]. (2023-10-02)[2025-04-08]. https://arxiv.org/abs/2310.01218v1.

作者简介

王 伟 男,1990年生. 博士,北京交通大学计算机科学与技术学院教授. 主要研究方向为计算机视觉、机器学习. 中国电子学会会员编号:E190029917M.
E-mail: wei.wang@bjtu.edu.cn



张静宜 女,2001年生. 北京交通大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: 24120305@bjtu.edu.cn

温玉辉 女,1990年生. 博士,北京交通大学计算机科学与技术学院副教授. 主要研究方向为计算机视觉、计算机图形学、机器学习.
E-mail: yhwen1@bjtu.edu.cn

魏云超 男,1986年生. 博士,北京交通大学计算机科学与技术学院教授. 主要研究方向为计算机视觉、机器学习.
E-mail: yunchao.wei@bjtu.edu.cn