

一种多模态跨领域的风格迁移方法

郑美君, 刘宏, 徐大宏, 李希

(湖南师范大学信息科学与工程学院, 湖南长沙 410081)

摘要: 风格迁移旨在将艺术风格应用于原始内容以生成创新作品。针对多模态风格迁移中类别关注不足及跨域适应性弱的问题, 提出一种基于 Vision Transformer (ViT) 和 Diffusion Transformer (DiT) 融合的多模态跨领域风格迁移方法。首先, 通过拼接 ViT 与 DiT 构建统一框架, 增强模型对文本或图像模态输入的协同处理能力, 利用 DiT 的深度扩展与标记扩充实现跨模态特征对齐, 解决多类别风格迁移中的模态偏差问题。其次, 改进 ViT 的自注意力机制, 在深层特征空间中解耦内容与风格特征, 并通过参数化融合模块实现跨领域风格适配。实验结果表明, 该方法在保留内容语义的同时, 显著提升了跨域风格迁移的多样性与艺术性, 定量指标与定性对比均优于现有方法, 验证了其在多模态跨域场景下的有效性。

关键词: 风格迁移; 多模态; 跨领域; 视觉 Transformer; 扩散 Transformer

DOI: 10.11907/rjdk.251343

扫描二维码阅读全文: 

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2026)008-0164-09

A Multi-Modal Cross-Domain Style Transfer Method

ZHENG Meijun, LIU Hong, XU Dahong, LI Xi

(College of Information Science & Engineering, Hunan Normal University, Changsha 410081, China)

Abstract: Style transfer aims to apply artistic style to original content to generate innovative works. This paper proposes a multi-modal cross domain style transfer method based on the fusion of Vision Transformer (ViT) and Diffusion Transformer (DiT) to address the issues of insufficient category attention and weak cross domain adaptability in multi-modal style transfer. Firstly, a unified framework is constructed by concatenating ViT and DiT to enhance the collaborative processing capability of the model for text or image modal inputs. DiT's deep and label extensions are utilized to achieve cross modal feature alignment and solve the problem of modal bias in multi class style transfer. Secondly, improve the self attention mechanism of ViT by decoupling content and style features in the deep feature space, and achieve cross domain style adaptation through a parameterized fusion module. The experiment shows that the proposed method significantly improves the diversity and artistry of cross domain style transfer while preserving content semantics. Both quantitative indicators and qualitative comparisons are superior to existing methods, verifying its effectiveness in multi-modal cross domain scenarios.

Key Words: style transfer; multi-modal; cross-domain; Vision Transformer; Diffusion Transformer

0 引言

风格迁移 (Style Transfer) 旨在将目标风格迁移至源内容图像, 生成兼具语义一致性和艺术表现力的视觉作品, 在数字艺术创作、影视特效等领域得到了广泛应用^[1-2]。然而, 现有方法仍面临两大核心挑战: 一是输入模态单一, 难以有效融合文本描述的抽象概念与图像具体特征^[3]; 二

是跨域迁移能力不足, 现有研究方法多聚焦于单风格^[4]或任意风格^[5]的研究, 对多风格以及不同风格间的跨域迁移^[6]探索不足, 难以平衡内容保真度与风格适配性。

近年来, 扩散模型 (Diffusion Models) 在高分辨率图像合成^[7]、语义分割^[8]、图像编辑^[9]等任务中展现出卓越性能, 为风格迁移提供了新思路。条件扩散模型虽能实现精准迁移, 但依赖成对训练数据^[10]; 而无条件模型, 如双扩散隐式桥, 虽放宽了数据要求, 但需为每个域独立训练模型,

收稿日期: 2025-06-23

作者简介: 郑美君 (1998-), 女, 湖南师范大学信息科学与工程学院硕士研究生, 研究方向为计算机视觉、风格迁移; 刘宏 (1963-), 男, 硕士, 湖南师范大学信息科学与工程学院教授、博士生导师, 研究方向为人工智能、智能家居; 徐大宏 (1971-), 男, 博士, 湖南师范大学信息科学与工程学院副教授、硕士生导师, 研究方向为机器视觉、大数据等; 李希 (1981-), 男, 博士, 湖南师范大学信息科学与工程学院讲师, 研究方向为计算机视觉、体育智能、区块链。本文通讯作者: 李希。

计算成本高昂^[11]。Diffusion CLIP通过CLIP编码器实现文本驱动迁移,但仍需微调模型以适配新风格^[12]。与此同时,Transformer架构的引入带来技术的突破性进展。DiT(Diffusion Transformer)用Transformer架构替换原扩散模型中的U-Net主干网络,证明了Transformer在跨模态对齐中的潜力^[13]。DINO-ViT(DINO-trained Vision Transformer)则通过自监督学习获得强语义表征,为内容风格特征解耦提供了新工具^[14]。特别地,Splice方法表明,ViT(Vision Transformer)的类别令牌可编码全局风格特征,而深层空间特征的自相似性可表征语义内容,从而为跨域风格迁移奠定了理论基础^[15]。

针对上述问题,本文提出一种多模态跨领域风格迁移方法DVST(Diffusion-Vision Transformers Style Transfer)。该方法的核心内容如下:首先,构建统一的文本—图像混合输入空间,将DiT与ViT首尾拼接,利用DiT跨模态对齐机制直接处理文本—图像的多种输入组合,突破单一图像的输入限制。其次,在ViT深层特征空间中利用DINO-ViT提取图像的风格表征与内容表征,并结合内容一致性损失与风格适配损失优化生成过程,实现跨域特征的有效融合。最后,设计零参数自适应策略,训练并记录模型使用统一参数权重处理未知风格—内容组合的最佳效果,使模型无需微调即可保持高质量的图像生成效果,有效提升了泛化能力。

1 相关工作

1.1 扩散模型

扩散模型因具有较强的高质量图像生成能力而变得流行^[16]。该生成模型通过马尔可夫链前向过程逐步加入高斯噪声,然后使用训练好的噪声估计模型,通过迭代去噪过程从潜在噪声中生成干净样本。具体来说,扩散模型从数据中抽取一个样本图片 x_0 ,再从 $t \in [1, \dots, T]$ 中随机选取一个时间 t ,并将 t 转化为对应向量,将 x_0 和转化后的向量传给扩散模型。扩散模型采样一个随机噪声 $\beta_t \in (0, 1)$ 叠加至 x_0 ,得到 x_t 。具体过程如下:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon \quad (1)$$

其中, $\varepsilon \sim N(0, I)$, $\alpha_t = 1 - \beta_t$, 且 $\bar{\alpha}_t = \prod_{i=0}^t \alpha_i$, 生成干净图像的反向采样过程如下:

$$x_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} \varepsilon_\theta(x_t, t) \right) + \sigma_t \varepsilon \quad (2)$$

神经网络用于估计噪声分量 $\varepsilon_\theta(x_t, t)$,其可被视为最高比例因子的分数函数。

1.2 Transformers 拓展

Transformers是一种基于自注意力机制的深度学习模型,由输入、编码器、解码器、输出部分组成^[17]。ViT(Vision Transformer)是Transformers在图像领域的应用,本质

上还是采用Transformers架构,但增加了一个图像预处理环节^[18]。DINO-ViT是基于ViT的一种拓展模型,已被证明能够学习强大且有意义的视觉表示,在图像检索、对象分割和拷贝检测等几项下游任务中取得了较好结果^[14]。DiT则用ViT替换了去噪扩散概率模型中原有的网络架构,构建了一种全新可扩展的扩散模型^[13]。

受益于扩散模型强大的高质量图像生成能力,基于Transformers标准架构,本文受DiT、DINO-ViT和Splice方法的启发构建了DVST方法。本文将DiT模块和ViT模块以Transformers标准架构进行首尾拼接,对齐文本图像等不同模态的信息,并支持多模态的内容或风格输入,从而实现更灵活的风格迁移;借助DINO-ViT的力量解耦内容和风格特征,再训练一个生成器将内容风格重新进行融合;学习Splice用全局[CLS]令牌提取所需风格特征,以键的自相似性表征语义内容的方法将内容和风格融合到深层ViT特征空间中,实现跨域风格迁移,最终生成高质量的图像输出。

2 DVST 方法

2.1 DVST 架构开发

本文建立的DVST模型架构如图1所示。与现有的风格迁移方法不同,本文方法主要包含两个关键阶段:①多模态输入的特征编码与对齐;②跨域风格迁移的内容风格融合与生成。

2.2 多模态输入

本文受DiT方法的启发,利用扩散模型逐步采集并加入合适噪声的方法成功将文本与图像不同模态的信息结合。如图1所示,文本被编码为与图像配对的标签 y ,再与经过Patchify模块处理得到的噪声图像,以及扩散模型所需的时间步长 t 一起送入DiT Block。DiT Block采用融合adaLN-Zero调制的Transformer Block,共堆叠 N 层。之后经过Layer Norm层,在每个通道(channel)上对每个样本进行标准化,使每个通道的均值为0,标准差为1。再经过Linear层(全连接层、密集层)和Reshape层执行线性变换,将输入数据映射到输出数据,得到 $32 \times 32 \times 4$ 的噪声。最后通过一个vae的decoder模型,将噪声转变成RGB通道 512×512 的中间图像($512 \times 512 \times 3$),便于观察与调整参数。结合文本可以抽象描述概念,图像可以具象表达信息的特点,依据风格迁移任务的内容和风格要素,将文本、图像、内容、风格拆分为4个独立元素。本文确定了4种内容—风格输入组合范式,如图2所示。首先是图像—图像,内容与风格皆由图像提供,模型将融合源内容图像的内容结构与源风格图像的艺术风格;其次是图像—文本,内容源于文本描述(文本定义语义内容,不限制具体结构布局),风格源于参考风格图像(提供具体视觉风格);再次是文本—图像,内容源于参考内容图像(提供具体视觉结构),风

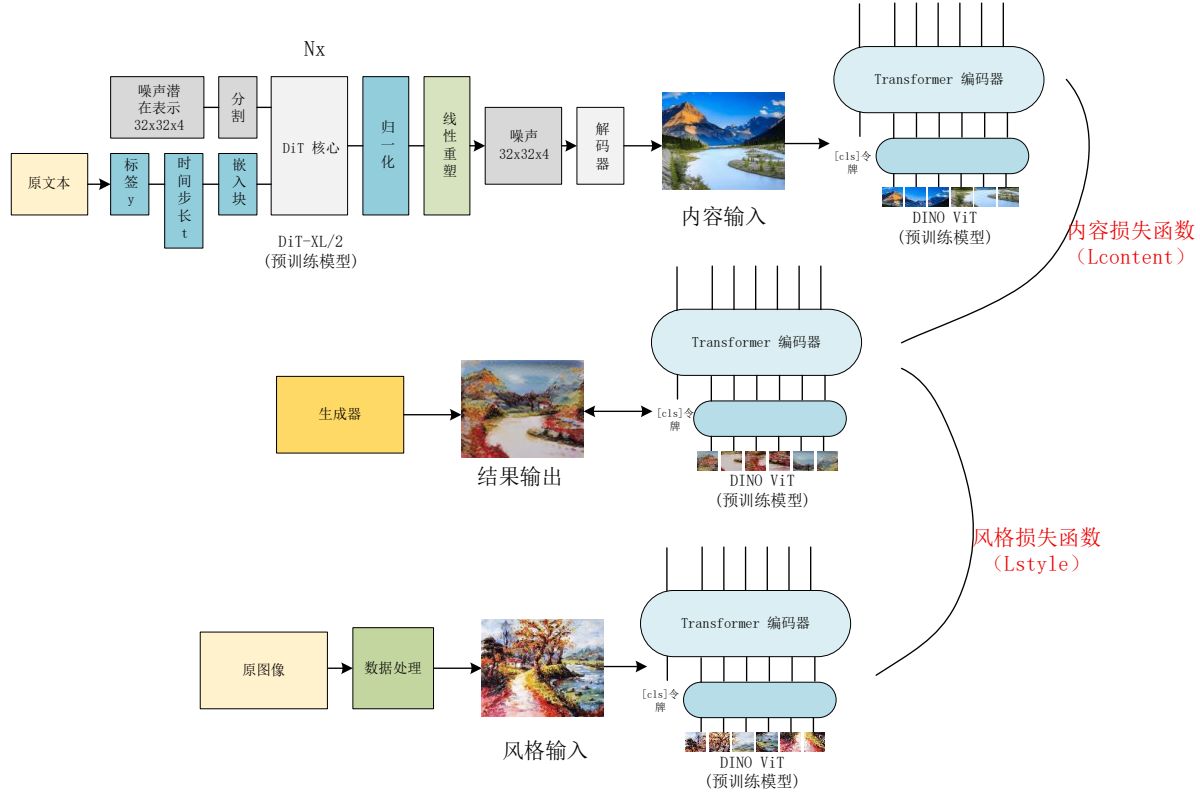


Fig. 1 DVST model architecture

图1 DVST模型架构

格源于文本描述的种类标签中所蕴含的风格特征(如不同种类鸟的外观);最后是文本—文本,内容和风格均源于文本描述。

2.3 跨领域融合

本文在 Splice 基础上进一步借助 DINO-ViT 的特征提取优势,除在任意两个语义相关对象(自然照片)之间拼接结构和语义外观外,增加了照片与其他艺术图片的跨领域融合尝试。为了让模型支持更多不同领域的图像数据,更好地学习不同领域间的信息差异,本文重写了与图像相关的数据加载、模型训练与数据保存等模块代码,使图像输入不再局限于一对结构—外观图像对,而是可以同时接受两个相关类别的多个图像输入,充分学习不同类别图像的信息表征。之后,为了适应不同领域信息的解耦与融合,本文通过反复实验逐步微调参数以确定最优 epoch,以防止不同领域图像间的过拟合现象。最后,通过修改损失函数,增添并修改相应参数的值,以进一步改善艺术风格主导而内容语义不足的问题。具体而言,本文用[CLS]令牌作为图像的风格表示,令牌集通过 L 个 Transformer 层传递,每个层由规范化层(Layer Normalization, LN)、多头自关注(Multi-Head Self-Attention, MSA)模块和多层感知机(Multi-Layer Perceptron, MLP)组成:

$$T^l = MLP(LN(MSA(LN(T^{l-1})) + T^{l-1})) + MSA(LN(T^{l-1})) + T^{l-1} \quad (3)$$

其中, $T^l(I) = [t_{cls}^l(I), t_1^l(I), \dots, t_n^l(I)]$ 是图像 I 第 l 层的输出令牌。

在每个 MSA 块中,规范化的令牌被线性地投影到查询、键和值中:

$$Q^l = T^{l-1} \cdot W_q^l, K^l = T^{l-1} \cdot W_k^l, V^l = T^{l-1} \cdot W_v^l \quad (4)$$

利用从 ViT 中提取的深层空间特征中的自相似性作为图像的内容表示:

$$S^l(I)_{ij} = \cos - \text{sim}(k_i^l(I), k_j^l(I)) \quad (5)$$

其中, $\cos - \text{sim}$ 是键之间的余弦相似度。自相似描述符的维数为 $S^l(I) \in R^{(n+1) \times (n+1)}$, n 为 patches 的个数。

最后,训练一个生成器来融合特征并生成最终图像,所使用的总损失函数为:

$$L_{\text{total}} = \alpha L_{\text{content}} + \beta L_{\text{style}} \quad (6)$$

其中, α 和 β 表示两项之间的相对权重,本文设 $\alpha = 0.1, \beta = 0.1$ 。

内容损失函数鼓励输出图像匹配内容图像的内容,其由在最深层注意力模块(Self-Sim)中所提取出键的自相似性差异进行定义:

$$L_{\text{content}} = \|S^L(I_c) - S^L(I_o)\|_F \quad (7)$$

其中, $S^L(I)$ 的定义见式(5)。

风格损失函数鼓励输出图像匹配风格图像的风格,其由输出图像与内容图像之间最深层的[CLS]令牌差异进行定义:

$$L_{\text{style}} = \|t_{[\text{CLS}]}^L(I_s) - t_{[\text{CLS}]}^L(I_o)\|_2 \quad (8)$$

其中, $t_{[\text{CLS}]}^L(\cdot) = t_{cls}^L$ 是从最深层提取的[CLS]令牌。

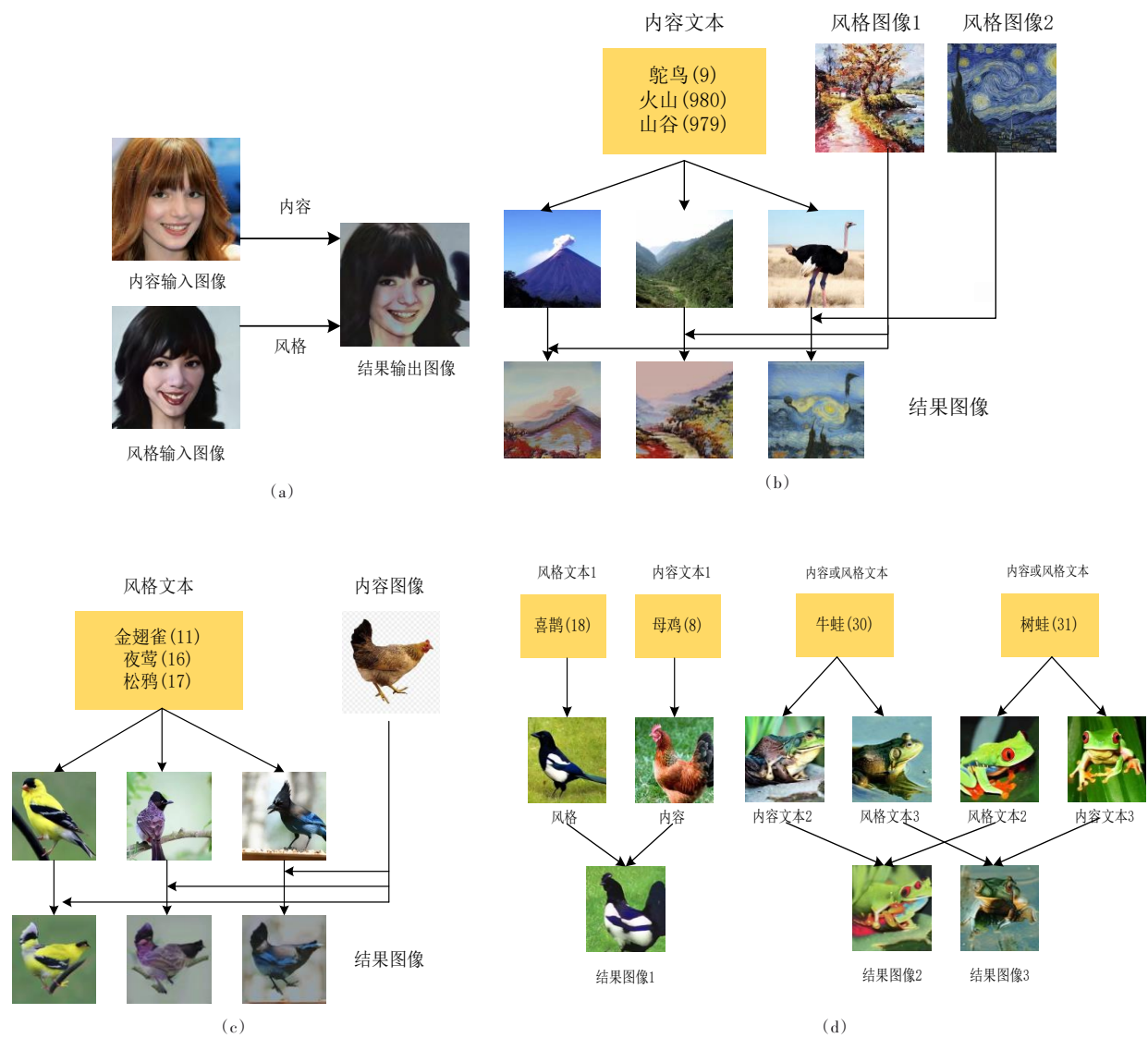


Fig. 2 Implementation methods of four multimodal input combinations

图 2 4种多模态输入组合的实现方式

3 实验与结果分析

3.1 实验设置

本文实验基于 NVIDIA V100 GPU, 网络使用 Pytorch 框架进行搭建, 利用 Python 进行编程。实验采用两个数据集: 一是包含 1 000 个类别的计算机视觉邻域常用的公开数据集 ImageNet, 二是自己构建的多类别、多风格数据集 Fusion(25 风格+75 内容)。本文使用颜色抖动、随机裁剪和水平/垂直翻转等增强技术优化训练过程。模型的总训练次数 n_epochs 值为 3 000, 梯度累积 gradient_accumulation_steps 值为 4, 图像记录频率参数 log_images_freq 设置为 10, 初始学习率设置为 0.002, 采用 Adam 优化算法进行参数更新。损失函数权重参数 lambda_entire_cls 与 lambda_entire_ssim 分别设置为 10 和 1。

3.2 训练损失与中间结果可视化

网络训练中的总损失、风格损失、内容损失曲线如图 3 所示。纵坐标为损失值, 横坐标为训练迭代次数。从总损失曲线看, 网络训练是有效收敛的, 且在 300 个 Epoch 内

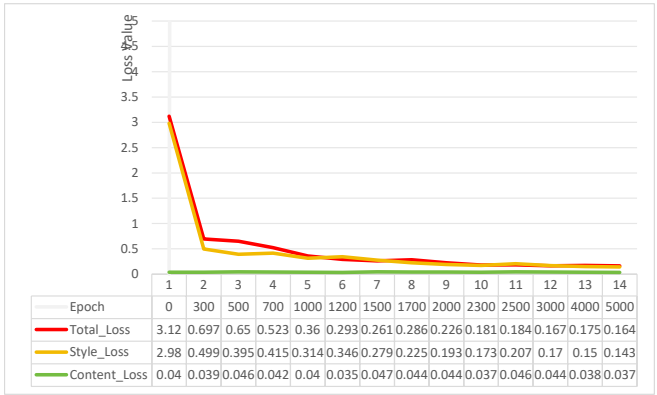


Fig. 3 Total loss, style loss, and content loss in network training

图 3 网络训练中的总损失、风格损失与内容损失

收敛速度较快,总损失趋于稳定,约为0.69。在3 000个Epoch时,总损失逐渐减小到0.16左右,风格损失曲线随总损失曲线同步减小至0.17左右,内容损失曲线一直保持在0.03~0.04左右,以上证明本文提出的网络架构达到了预期效果。

网络训练中不同Epoch下生成的中间图像结果如图4所示。对于每个示例,从左到右依次为:内容表示、风格表示和Epoch为300、500、1 000、1 500、2 000、3 000的中间

图像结果。可以看到,当Epoch设置为3 000,所有示例都已展现出优异的合成效果。针对两组示例:同内容不同风格,同风格不同内容,本文方法都成功在保留原内容的基础上实现了有效的风格迁移。如第一、二行山水画中的山、水、云等元素,第三行山的清晰前景和模糊后景等内容,在各个阶段结果中都被有效地保留了下来。而两幅不同风景油画中的颜色、笔触、纹理等风格,也被有效地迁移到各自的原参考内容图像上。



Fig. 4 Generated intermediate image results with different epochs

图4 不同Epoch下生成的中间图像结果

3.3 Fusion数据集扩展

本文以自监督方式训练的预训练ViT模型(DINO-ViT)作为特征提取器^[14]。在8×8补丁配置(ViT-B/8)中使用了12层预训练模型,该模型可从GitHub官方下载。之后本文基于DINO-ViT语义规则,挑选语义明确、边界清晰的图像作为内容或风格图像。图像格式包括jpg、jpeg、png 3种,图像分辨率范围为256~2 000 px。图像来源为网上各种公共数据集、开源平台、人工智能生图工具等,如MS-COCO数据集、AFHQ数据集、WikiArt数据集、Flickr、Pixabay、豆包等。数据集规模由原来的40类(20风格+20内容)扩展至100类(25风格+75内容)。其中,风格包括水墨、线稿、炭笔涂鸦、水溶彩铅、蜡笔、中西写意、新工笔、水彩画、版画、国风、绘本、插画、简笔画、油画、儿童画、荧光色绘画、CG厚涂、伦勃朗、素描、暗黑童话、动漫、复古动漫、印象派、野兽派、黑白水墨,内容涵盖艺术和摄影中广泛应用的人物、动物、植物、静物四大类。如图5所示,在内容相同风格不同、内容不同风格相同、内容不同风格不同的3组实验中,本文方法成功应用于各种具有挑战性的跨域风格迁移任务中,验证了本文算法的有效性。

对于分别以40类与100类Fusion数据集训练的DVST

模型,本文通过深度Gram矩阵量化纹理特征匹配度(风格契合度),并通过结构相似度(Structural Similarity, SSIM)量化结构相似性(内容保真度)。在40类规模与100类规模的实验结果中,分别抽取50个样本进行估算。表1的估算数据显示,不管从风格契合度还是内容保真度,模型效果都有所提升,验证了本文模型性能的稳定性。

3.4 评价指标

为了更好地衡量DVST方法改进后的性能,本文同时使用了定性和定量指标。考虑到DVST方法包含多模态文生图生成与跨域风格迁移两个核心任务,评估指标的设计分别针对以下两方面:

(1)多模态生成质量评估。FID score是一种用于评估生成图像质量的度量标准,专门用于评估模型的图像生成质量^[19]。计算公式如下:

$$FID(p, q) = \|\mu_p - \mu_q\|_2^2 + T_r(C_p + C_q - 2\sqrt{C_p C_q}) \quad (9)$$

其中, p 表示真实图像的分布, q 表示生成图像的分布。 μ_p 和 μ_q 分别表示两个分布特征向量的均值, C_p 和 C_q 分别表示两个分布特征向量的协方差矩阵, T_r 表示矩阵的迹运算。 $\|\cdot\|_2$ 表示欧几里得范数,常用于计算向量长度。

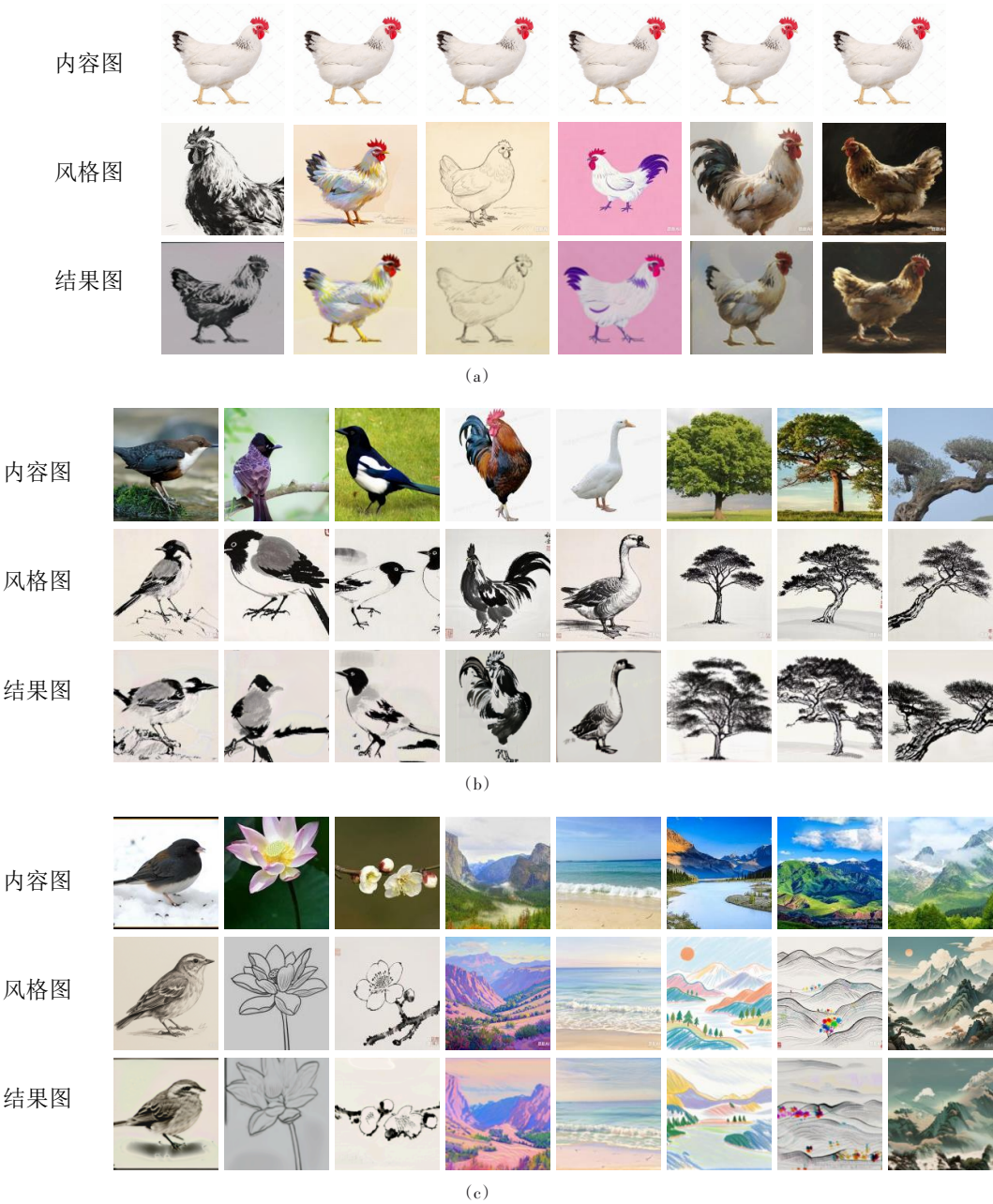


Fig. 5 Experimental results of the expansion of the Fusion dataset
图 5 Fusion数据集扩展实验结果

Table 1 Comparison of styles and content at different scales
表 1 不同规模下的风格与内容比较

类别规模	风格契合度	内容保真度
40类	0.742 3	0.745 4
100类	0.785 6	0.791 8

此外, SFID (Sliced Frechet Inception Distance) 用于衡量生成图像的质量和真实性^[20]。IS (Inception Score) 是用于评估生成模型生成图像质量和多样性的一个经典指标^[21]。

(2) 跨域风格迁移质量评估。Human Perceptual Evaluation 是指人类级别的表现、人类级别的精度等, 范围涵盖图像生成、风格迁移、物体检测等各个方面^[22]。本文通过用户主观评分, 从人类感知角度评估生成图像在内容保真

度与风格一致性方面的表现, 这对于验证多类别跨域迁移能力尤为重要。

3.5 方法定性比较

将本文方法与 Splice、Zecon、DEADiff 3 种风格迁移的经典方法进行了比较。其中, Splice 是一种经典的风格迁移方法, 用于提取图像的语义特征, 可实现图像语义外观之间的转换^[15]。Zecon 是一种经典的对比损失方法, 可通过文本引导提取图像风格特征, 实现零样本条件下的图像风格迁移^[23]。DEADiff 是一种经典的高效扩散模型, 用于解耦图像表示以提取图像的风格特征, 从而实现图像风格的高效转换^[24]。

(1) 与图一图风格迁移模型 Splice 的比较。定性比较

的实验结果如图6所示,从上到下依次为4组数据:人脸和人脸跨艺术风格领域的风格转换、梵高星空风景图和动物的跨学科领域风格转换、风景照片和梵高星空跨艺术风格领域的风格转换、风景照片和油画风景图跨艺术风格领域的风格转换。从左到右依次为:风格表示、内容表示、Splice模型结果以及本文模型结果。本文模型生成结果相比Splice模型,在相同的迭代次数下,不管是人脸五官内容的保留,还是动物语义内容的保留,亦或是风景图片的布局保留,都很好地学习到对应的内容信息,且展现出有效的跨领域风格迁移能力。反之,对于Splice生成的图像,第一行眼睛鼻子的内容细节丢失,第二行鸵鸟身体的风格元素丢失,第三行的河流中有明显的伪影,第四行的图片整体都很模糊。

(2)与文字引导艺术风格迁移模型ZeCon的比较。定性比较的实验结果如图7所示,从上到下依次为2组数据:河流风景图的Crayon(蜡笔)风格迁移,河流风景图的Snowy(雪景)风格迁移。从左到右依次为:内容表示、Zecon文字风格表示、ZeCon模型结果、本文模型结果以及与Zecon文字风格表示同类型的图像风格表示。本文模型生成的图像在内容保留方面,保留了河流的走向、两岸草丛的布局等空间结构信息;在风格迁移方面,第一行充分体现了Crayon风格的蜡笔笔触质感,第二行真实还原了Snowy风格的积雪覆盖效果,保持了内容与风格一致的高保真度。而ZeCon模型的结果都只具备了局部的风格信

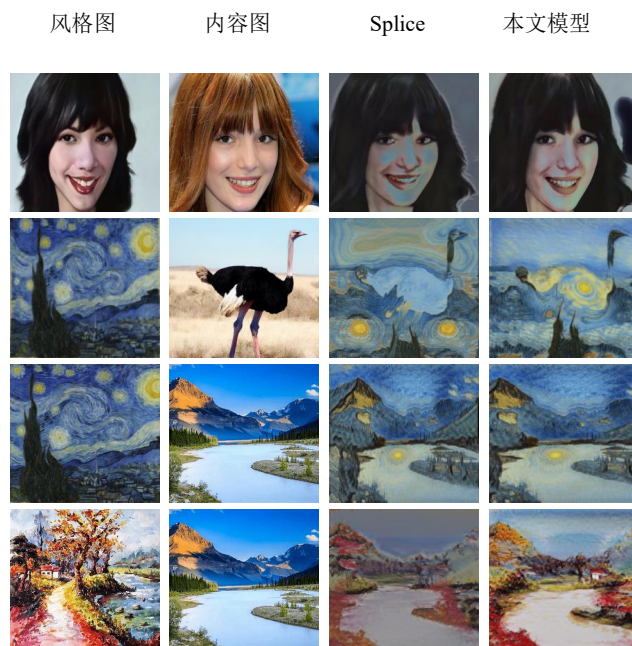


Fig. 6 Qualitative comparison between DVST and the classical model

Splice

图6 DVST与经典模型Splice的定性比较

息,在Crayon风格中,河流左下区域缺失风格元素的作用;在Snowy风格中,整个河流和旁边的草丛风格化程度不足,有时还受提示词的影响出现了伪影。



Fig. 7 Qualitative comparison between DVST and the classical model ZeCon

图7 DVST与经典模型ZeCon的定性比较

(3)与具有解纠缠表示的高效风格化扩散模型DEADiff的比较。定性比较的实验结果如图8所示,从上到下依次为3组数据:第一组以梵高艺术画作为风格参考、Ostrich(鸵鸟)文本作为内容参考,第二组以梵高艺术画作为风格参考、Volcano(火山)文本作为内容参考,第三组以油画风景图作为风格参考、Volcano(火山)文本作为内容参考。从左到右依次为:风格表示、内容表示、DEADiff模型结果以及本文模型结果。本文模型相比DEADiff模型更贴近保留了原参考内容和风格。如图8所示,在第一、二行中,本文模型在风格保留方面,保留了原参考风格的蓝、黑、黄三色,并体现了黑色椰子树的抽象风格;而

DEADiff模型只保留了蓝、黄两种颜色,且黑色椰子树的风格特点、原风格的抽象艺术特点都没有体现。在内容保留上,本文也更贴近原文本描述,如图8的第一行,本文保留了完整的鸵鸟语义内容,而DEADiff模型只保留了鸵鸟的头部语义。但对于第三行的油画风格与火山这一目标,DEADiff模型与本文模型均取得了良好效果。

3.6 方法定量比较

在定量指标中,FID与SFID以最小化为优化目标,IS以最大化为优化目标。借鉴DiT方法,本文多模态风格迁移中的文字图像语义信息对齐能力以在class-conditional 256×256 ImageNet生成基准上的得分情况(见表2)与

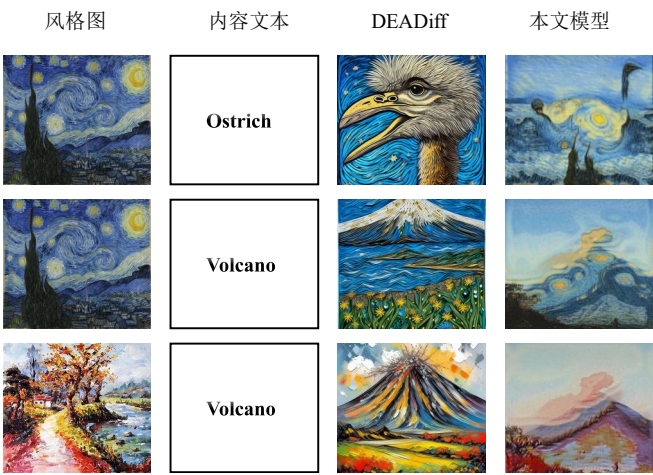


Fig. 8 Qualitative comparison between DVST and the classic model DEADiff

图8 DVST与经典模型DEADiff的定性比较

class-conditional 512×512 ImageNet生成基准上的得分情况(见表3)来衡量^[13]。可以看出,无论是在256×256还是512×512规模上,本文算法均展现出更好的性能,超越了所有对照的经典算法。

Table 2 Benchmark results of class-conditional image generation on ImageNet at 256×256 resolution

表2 ImageNet 256×256分辨率下类标签条件图像生成的基准测试结果

模型	FID 指标	sFID 指标	IS 指标
BigGAN-deep	6.95	7.36	171.40
StyleGAN-XL	2.30	4.02	265.12
ADM	10.94	6.02	100.98
ADM-U	7.49	5.13	127.49
Adm-g	4.59	5.25	186.70
Ldm-8	15.51	-	79.03
Ldm-8-g	7.76	-	209.52
Ldm-4	10.56	-	103.49
本文模型	2.27	4.60	278.24

Table 3 Benchmark results of class-conditional image generation on ImageNet at 512×512 resolution

表3 ImageNet 512×512分辨率下类标签条件图像生成的基准测试结果

模型	FID 指标	sFID 指标	IS 指标
BigGAN-deep	8.43	8.13	177.90
StyleGAN-XL	2.41	4.06	267.75
ADM	23.24	10.19	58.06
ADM-U	9.96	5.62	121.78
Adm-g	7.72	6.57	172.71
本文模型	3.04	5.02	240.82

本文通过用户调查,从人类主观感知角度评估跨领域图像风格迁移的生成质量。根据惯例,本文采用文献[25]中提出的两种强制选择(Two-Alternative Forced Choice, 2AFC)协议。每位参与者会看到2张参考图像(输入内容图像(img1)、输入风格图像(img2))以及4个备选结果(本文结果和其他几个模型的结果)。参与者被问到:“哪张图

片最能展示img1的语义内容和img2的颜色风格?”本文总共使用100张图片进行统计,大致收集了100个用户判断。表4显示了票数百分比,本文方法在ImageNet数据集、Fusion数据集的表现均优于其他模型。

Table 4 User surveys on cross-domain style migration generated image quality

表4 跨域风格迁移生成图像质量的用户调查 (%)				
数据集	Zecon	Splice	DEADiff	本文模型
ImageNet数据集	21 ± 5.0	24 ± 5.0	26 ± 6.0	29 ± 3.0
Fusion数据集	21 ± 6.0	24 ± 7.0	25 ± 5.0	30 ± 2.0

4 结语

针对多模态风格迁移中类别关注不足及跨域适应性问题,本文提出的DVST方法能够在保留原参考内容和风格的同时,生成高质量的创新作品。与现有方法相比,本文通过构建文本—图像混合输入空间,并借助DiT的跨模态对齐机制,实现了对文本—图像多种输入组合的处理,有效突破了传统方法单一图像输入的限制。同时,基于DINO-ViT在ViT深层特征空间中的解耦提取风格与内容表征,并融合Splice方法提出的内容一致性损失与风格适配损失进行联合优化,促进了跨域特征的有效融合。在此基础上,设计零参数自适应策略,使模型能够以统一参数权重适应未知的风格—内容组合,无需额外微调即可维持高质量的图像生成。实验结果表明,该方法在多项测试中均具有较好表现,为多模态跨域图像风格迁移任务提供了一条可行的技术路径。未来可通过引入更多样化的训练数据(如更多类别的内容数据集与风格数据集),并采用自监督或无监督学习方法继续增强算法的泛化能力,进一步提升模型表现。

参考文献:

[1] LI C, SHI J P, JIANG L. Image multi-style transfer algorithm in deep learning[J]. Software Guide, 2021, 20(7): 171-176.
李超,石剑平,姜麟. 深度学习中的图像多风格迁移算法[J]. 软件导刊, 2021, 20(7): 171-176.

[2] LIAN L, TIAN Q C, TAN R, et al. Research progress on image style transfer based on neural networks [J]. Computer Engineering and Applications, 2024, 60(9): 30-47.
廉露,田启川,谭润,等. 基于神经网络的图像风格迁移研究进展[J]. 计算机工程与应用, 2024, 60(9): 30-47.

[3] PAN S Y, ZHAO Z P, YANG Q X, et al. Artistic style transfer based on ViT semantic guidance and structure-aware enhancement [J]. Chinese Journal of Computers, 2025, 48(9): 2131-2158.
潘书煜,赵征鹏,阳秋霞,等. 基于ViT语义指导与结构感知增强的艺术风格迁移[J]. 计算机学报, 2025, 48(9): 2131-2158.

[4] LI W W, FU B, WANG H F, et al. Image ink painting style transfer method based on data enhanced recurrent generative adversarial network [J]. Journal of Jilin University (Science Edition), 2025, 63(3): 804-814.
李伟伟,傅博,王贺霏,等. 基于数据增强循环生成对抗网络的图像水墨画风格迁移方法[J]. 吉林大学学报(理学版), 2025, 63(3):

- 804–814.
- [5] WANG S S, LI W S. A review of research on arbitrary image style transfer based on neural networks and attention [J]. *Software Engineering*, 2025, 28(2): 27–31.
王树声, 李文书. 基于神经网络与注意力的任意图像风格迁移研究综述[J]. *软件工程*, 2025, 28(2): 27–31.
- [6] ZHOU Y, CHEN Z C, HUANG H. Deformable one-shot face stylization via DINO semantic guidance[DB/OL]. <https://arxiv.org/pdf/2403.00459>.
- [7] ROBIN R, ANDREAS B, DOMINIK L, et al. High-resolution image synthesis with latent diffusion models[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 10674–10685.
- [8] DMITRY B, ANDREY V, IVAN R, et al. Label-efficient semantic segmentation with diffusion models [DB/OL]. <https://arxiv.org/pdf/2112.03126>.
- [9] OMRI A, DANI L, OHAD F. Blended diffusion for text-driven editing of natural images[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 18187–18197.
- [10] SAHARIA C, CHAN W, CHANG H, et al. Palette: image-to-image diffusion models [C]// *ACM SIGGRAPH 2022 Conference Proceedings*, 2022: 1–10.
- [11] SU X, SONG J, MENG C, et al. Dual diffusion implicit bridges for image-to-image translation[DB/OL]. <https://arxiv.org/pdf/2203.08382>.
- [12] GWANGHYUN K, TAESUNG K, JONG C. DiffusionCLIP: text-guided diffusion models for robust image manipulation [C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 2426–2435.
- [13] PEEBLES W, XIE S. Scalable diffusion models with transformers [C]// *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023: 4172–4182.
- [14] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers [C]// *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 9630–9640.
- [15] TUMANYAN N, BAR-TAL O, BAGON S, et al. Splicing ViT features for semantic appearance transfer [C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 10738–10747.
- [16] JONATHAN H, AJAY J, PIETER A. Denoising diffusion probabilistic models [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 6840–6851.
- [17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.
- [18] ALEXEY D, LUCAS B, ALEXANDER K, et al. An image is worth 16x16 words: transformers for image recognition at scale [DB/OL]. <https://arxiv.org/pdf/2010.11929>.
- [19] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [C]// *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6629–6640.
- [20] CHARLIE N, JACOB M, SANDER D, et al. Generating images with sparse representations [DB/OL]. <https://arxiv.org/pdf/2103.03841>.
- [21] TIM S, IAN G, WOJCIECH Z, et al. Improved techniques for training GANs [C]// *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016: 2234–2242.
- [22] JING Y, YANG Y, FENG Z, et al. Neural style transfer: a review [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2019, 26(11): 3365–3385.
- [23] 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*, 2023: 22816–22825.
- [24] QI T, FANG S, WU Y, et al. DEADiff: an efficient stylization diffusion model with disentangled representations [C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 8693–8702.
- [25] PARK T, ZHU J Y, WANG O, et al. Swapping autoencoder for deep image manipulation [DB/OL]. <https://arxiv.org/pdf/2007.00653>.

(责任编辑:黄健)