智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于跨模态注意力的多阶段图像风格迁移

仇佳庆, 窦立云, 王进

引用本文:
仇佳庆, 窦立云, 王进. 基于跨模态注意力的多阶段图像风格迁移[J]. 智能系统学报, 2026, 21(3): 751-762.
QIU Jiaqing, DOU Liyun, WANG Jin. Multistage image style transfer based on cross-modal attention[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 751-762.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202508011

DOI: 10.11992/tis.202508011
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260324.1659.005

基于跨模态注意力的多阶段图像风格迁移

仇佳庆, 窦立云, 王进
(南通大学 人工智能与计算机学院, 江苏 南通 226019)

摘　要: 图像风格迁移 (image style transfer, IST) 的核心是融合图像内容与目标艺术风格, 生成兼具语义合理性与视觉表现力的图像, 广泛应用于艺术创作、个性化图像编辑等领域。现有方法处理高分辨率图像、复杂纹理及跨模态引导任务时, 存在计算效率低、风格可控性弱、细节丢失、风格与内容脱节等问题。为此, 本文提出基于潜在扩散模型的多阶段风格迁移框架 CAST-Diff(cross-modal attention and style-adaptive diffusion framework), 采用解耦式协同设计, 结合跨模态语义引导、自适应区域风格调节与潜空间扩散细化, 通过跨模态注意力对齐图文特征, 利用自适应模块控制区域风格强度, 利用潜扩散模型完成去噪与细节重建。在 COCO、Flickr30k 数据集上与 StyleGAN-Diffusion、ControlNet 等主流方法对比表明, CAST-Diff 在风格一致性、细节保真度及视觉自然度上更优, 能在复杂场景下保留图像结构与精细纹理, 实现自然逼真的风格迁移, 同时提升计算效率与泛化能力, 为文本引导的高精度风格迁移提供可行方案。

关键词: 图像风格迁移; 跨模态注意力; 自适应风格调节; 潜在扩散模型; 图像生成; 文本引导图像生成; 多模态生成; 艺术风格转换
中图分类号: TP391; TH212　　**文献标志码:** A　　**文章编号:** 1673−4785(2026)03−0751−12

中文引用格式: 仇佳庆, 窦立云, 王进. 基于跨模态注意力的多阶段图像风格迁移 [J]. 智能系统学报, 2026, 21(3): 751–762.
英文引用格式: QIU Jiaqing, DOU Liyun, WANG Jin. Multistage image style transfer based on cross-modal attention[J]. CAAI transactions on intelligent systems, 2026, 21(3): 751–762.

Multistage image style transfer based on cross-modal attention

QIU Jiaqing, DOU Liyun, WANG Jin
(School of Artificial Intelligence and Computer Science, Nantong University, Nantong 226019, China)

Abstract: Image style transfer (IST) aims to fuse image content with a target artistic style to generate semantically rational and visually expressive images, and it has been widely applied in art creation, personalized image editing, and other fields. Existing methods suffer from low computational efficiency, weak style controllability, loss of details, and disconnection between style and content when handling high-resolution images, complex textures, and cross-modal guided tasks. To address these issues, this paper proposes CAST-Diff, a multistage image style transfer framework built on latent diffusion models. CAST-Diff follows a decoupled collaborative design that combines cross-modal semantic guidance, adaptive regional style regulation, and latent space diffusion refinement so that text and image features can be aligned through cross-modal attention, regional style strength can be adjusted by the adaptive module, and denoising together with detail reconstruction can be completed by the latent diffusion model. Experimental comparisons with mainstream methods such as StyleGAN-Diffusion and ControlNet on the COCO and Flickr30k datasets show that CAST-Diff performs better in style consistency, detail fidelity, and visual naturalness while preserving image structure and fine textures in complex scenes, producing more natural and realistic style transfer results, and improving computational efficiency as well as generalization ability. These advantages make CAST-Diff a practical solution for text-guided high-precision style transfer.

Keywords: image style transfer; cross-modal attention; adaptive style modulation; latent diffusion model; image generation; text-guided image generation; multimodal generation; artistic style transfer

收稿日期: 2025−08−09.　**网络出版日期:** 2026−03−25.
基金项目: 南通市自然科学基金面上项目 (JC2025090).
通信作者: 窦立云. E-mail: Liyun_dou@163.com.

图像风格迁移 (image style transfer, IST) 是将图像内容与目标艺术风格进行融合并生成具有艺

术表现力结果的一类方法[1-4]，该方法目前已经应用于艺术创作、增强现实以及个性化图像编辑等任务，而基于卷积神经网络的风格迁移方法持续推动了这一方向的发展，也使相关技术在生成质量、运行效率以及表现力等方面不断提升。

现有方法在高分辨率图像、复杂纹理场景以及跨模态语义引导任务中仍然存在一些问题，计算效率偏低这一情况使其较难适应高分辨率生成需求，风格强度控制不够灵活则容易带来细节缺失或风格表达不足的问题，内容与风格之间的语义一致性建模也较为薄弱，因此很难契合复杂跨模态控制的标准。

针对上述问题，本文提出一种融合跨模态注意力、自适应风格调节与扩散模型的多阶段风格迁移框架。该框架主要包含三部分：基于 CLIP 的跨模态注意力机制，增强风格迁移的语义一致性；区域感知自适应调节策略，动态平衡细节保留与风格渲染效果；多阶段潜在扩散模型，通过由粗到精的逐步去噪实现高质量风格重构，有效提升生成结果的自然度与稳定性。

1　相关工作

1.1　图像风格迁移

图像风格迁移[5-8]已成为图像生成与视觉创作中的核心任务。早期方法依赖手工特征匹配，难以捕捉风格与内容间的深层语义关联。随着深度学习的发展，基于卷积神经网络 (convolutional neural network, CNN) 和生成模型的风格迁移方法在效果与效率方面均取得显著进展。Gatys 等[9]提出的神经风格迁移 (neural style transfer, NST) 方法通过最小化内容与风格损失实现优化生成，但其计算成本高。后续工作如 StyTr^2[7] 和 MLA-Net[1] 引入前馈式网络和注意力机制，相较传统优化方法显著加快了生成速度。

为提升风格可控性与内容保真，Huang 等[10]提出 AdaIN，实现任意风格快速迁移；Li 等[11]的 WCT(whitening and coloring transform) 方法通过特征空间变换缓解结构扭曲问题[8,12]。注意力机制的引入进一步增强了模型对图像区域的感知能力并提升了风格一致性的建模效果，AdaAttN 和多尺度自注意力方法[12-15]就是这方面的代表。Transformer 架构因具备较强的全局建模能力而被广泛用于风格迁移任务，StyTr^2 以及基于自注意力的多层融合方法[7]都在长程依赖建模与风格表达能力提升方面取得了较好效果。

近年来，任意风格迁移与跨模态风格迁移逐渐成为该领域关注的重点，CAST(cross-modal attention-based style transfer) 通过引入对比学习增强了模型的风格泛化能力，IEST(image enhanced style transfer) 则结合统计先验提升了方法的稳定性[15-17]。文本引导的风格迁移也在这一过程中逐步发展起来，Text2Style 与 ArtFlow 等方法[18-24]借助图文联合建模提升了语义一致性，并且为风格迁移提供了更多可调控的维度。

尽管已有方法在迁移质量、运行速度以及可控性方面已经取得一定进展，但这些方法在高分辨率图像处理、细粒度区域调节以及多模态一致性建模方面仍然存在不足，相关研究仍需继续推进更高效可控的迁移机制。

1.2　文本引导图像合成

文本引导图像合成 (text-to-image generation, T2I) 是根据自然语言描述生成语义一致并且视觉真实图像的任务[20]，现已用于内容生成、艺术创作以及人机交互等领域。T2I 作为跨模态生成任务发展较快，而其中的关键问题是复杂文本语义的建模，并且要使文本语义与图像内容实现有效对齐。

初期方法大多基于生成对抗网络 (generative adversarial network, GAN)，通过构建条件生成结构实现从文本到图像的映射，StackGAN 和 AttnGAN 引入分层生成与注意力机制增强文本到图像一致性[3,17,19]。但 GAN 训练不稳定[1-2]、控制性弱，使其在复杂语义场景中的扩展能力受到限制。

扩散模型的兴起[13,20-21]显著推动了 T2I 的进展。DALL·E 2、Imagen 与 Stable Diffusion 等模型通过逐步去噪生成高质量图像，兼顾细节、结构与语义一致性。LDM(latent diffusion model)[15,19-20,22]进一步将生成过程映射至潜空间，兼顾效率与图像保真度，成为当前主流方案。

语义编码器是 T2I 的关键组件[20-22]。CLIP (contrastive language-image pre-training) 通过大规模图文对比学习构建共享表示空间，为文本驱动生成提供强有力语义引导。基于 CLIP 的方法不断涌现，StyleCLIP[23] 和 Style GAN-NADA(CLIP-guided domain adaptation of image generators)[24] 在潜在空间中实现风格可控变换，VQGAN-CLIP(vector quantized GAN + contrastive language-image pre-training) 和 CLIPstyler 则结合内容与风格损失精细引导图像生成[25-28]，显著提升了表达力与一致性。

为实现更强的可控性与结构表达，研究者将结构先验作为辅助条件引入文本引导图像合成任务，ControlNet、T2I-Adapter 和 Classifier-Free Diffusion 等方法[20,23,25]通过融合视觉提示增强了图像

细节建模能力, 并且提升了区域控制效果, 融合注意力图以及多模态提示等机制也进一步提高了语义建模精度, 使生成过程更稳定。

文本引导图像合成正朝着高分辨率、多模态融合以及细粒度控制方向发展, 而如何实现更高保真度、更强泛化性以及更精细的语义风格建模, 仍是当前需要继续解决的关键问题。

2　实验方法

为使风格迁移模型在高分辨率图像合成的效果与效率有所提高, 本文研发了一种新型的多阶段风格迁移框架, 该框架涵盖了跨模态注意力、自适应风格迁移以及扩散型优化相关模块, 借助逐步地优化图像, 该框架把目标风格信息切实有效地融合到图像中, 同时把图像原有的结构及细节完整保留, 最终产出符合语义标准的高质量风格化图像。

跨模态注意力模块会把图像与文本描述一起映射到共享语义空间当中, 以此达成对风格的精准引导; 自适应风格迁移模块会根据图像不同区域的特征, 动态调整风格迁移的程度, 很好地避免了传统方法中容易产生的风格迁移过度或细节丧失等问题; 扩散型优化模块以逐步去噪的迭代过程来推进, 让图像细节质量以及风格一致性显著增强。

为解决传统方法在高分辨率图像处理中的计算难题, 本文研发了分阶段优化策略, 能一步步融合风格信息, 增强细节的表达能力; 通过在不同阶段对风格迁移和细节保留任务加以分离, 这个框架显著降低了计算复杂度, 也保证了生成图像的质量以及细节的准确情况。其具体优势是在 3 个方面体现: 1) 自适应风格控制机制可使不同区域风格调节达到更精细的程度; 2) 跨模态注意力模块把图像和文本信息有效结合在一起, 使风格迁移的灵活性和语义一致性得到提升; 3) 扩散模型的逐步推进优化, 避免了传统单阶段风格迁移产生失真问题。

基于这些创新, 本文的框架在多个基准数据集之上展现出良好的风格迁移效果, 生成的图像不仅细致而且高度相符, 并且在计算效率上与传统方法对比也有显著提升。

2.1　网络体系架构

所提出的风格迁移网络以内容图像 $\boldsymbol{I}_{\mathrm{C}}$ 与风格文本描述 $\boldsymbol{T}_{\mathrm{S}}$ 为输入, 并采用多阶段结构逐步优化风格融合过程, 最终输出风格化图像 $\boldsymbol{I}_{\mathrm{cs}}$。整个网络架构借助多层次特征提取、跨模态注意力机制以及潜空间扩散模型, 对图像风格迁移效果进行逐步优化。网络由图像编码器、文本编码器、跨模态注意力模块、潜在空间扩散模型以及解码器模块组成, 各模块在协同作用下实现内容图像结构与目标风格的准确结合, 并生成高质量风格化图像。

如图 1, 首先内容图像 $\boldsymbol{I}_{\mathrm{C}}$ 和风格文本描述 $\boldsymbol{T}_{\mathrm{S}}$ 分别输入到图像编码器和文本编码器开展特征提取。图像编码器采用 VGG 网络 (visual geometry group network) 提取内容图像特征 $\boldsymbol{f}_{\mathrm{C}}$, 文本编码器基于 CLIP 来开展风格文本语义特征 $\boldsymbol{f}_{\mathrm{S}}$ 的提取:

$$\boldsymbol{f}_{\mathrm{C}}=\boldsymbol{\varepsilon}_{\mathrm{img}}(\boldsymbol{I}_{\mathrm{C}}),\boldsymbol{f}_{\mathrm{S}}=\boldsymbol{\varepsilon}_{\mathrm{text}}(\boldsymbol{T}_{\mathrm{S}}) \tag{1}$$

式中: $\boldsymbol{I}_{\mathrm{C}}$ 表示内容图像; $\boldsymbol{T}_{\mathrm{S}}$ 表示风格文本描述; $\boldsymbol{\varepsilon}_{\mathrm{img}}$ 表示图像编码器, 并用于提取图像特征; $\boldsymbol{\varepsilon}_{\mathrm{text}}$ 表示文本编码器, 并用于提取风格文本的语义特征。

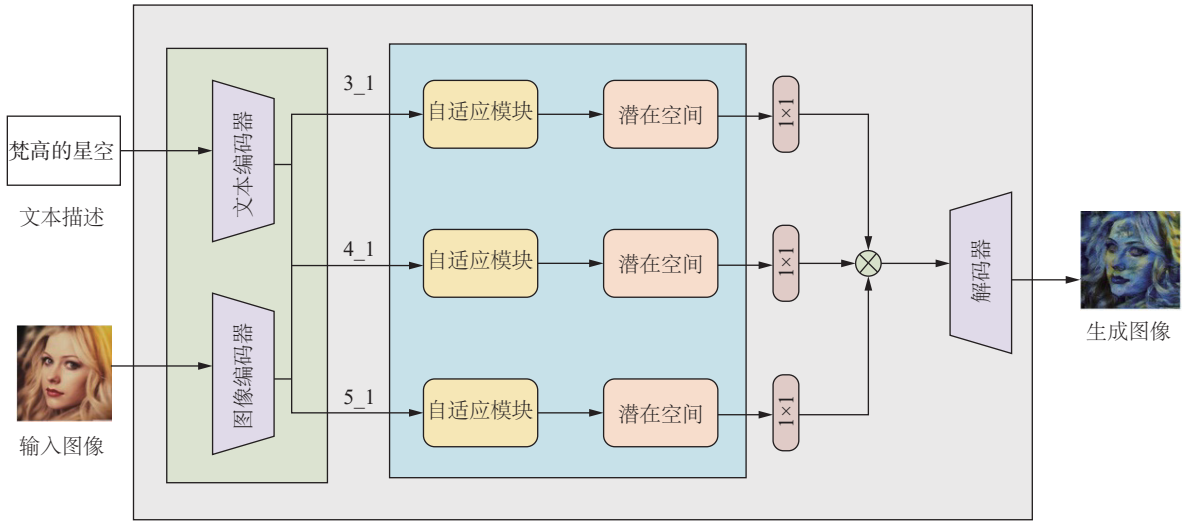

图 1　CAST-Diff 多阶段风格迁移框架整体架构

Fig. 1　Overall architecture of the CAST-Diff multi-stage style transfer framework

在多层次特征提取阶段，内容图像$\boldsymbol{I}_{\mathrm{C}}$和风格文本描述$\boldsymbol{T}_{\mathrm{S}}$输入多个特征提取层。这些层通过卷积操作提取图像的低级特征和高级特征。通过图像编码器的不同层，即 ReLU_{3_1}、ReLU_{4_1}、ReLU_{5_1}，分别提取内容图像$\boldsymbol{I}_{\mathrm{C}}$和风格文本描述$\boldsymbol{T}_{\mathrm{S}}$的特征图：

$$\boldsymbol{F}_r^{3\_1}=V_{\mathrm{encoder}}\left(\boldsymbol{I}_{\mathrm{C}},\mathrm{ReLU}_{3\_1}\right) \tag{2}$$

$$\boldsymbol{F}_r^{4\_1}=V_{\mathrm{encoder}}\left(\boldsymbol{I}_{\mathrm{C}},\mathrm{ReLU}_{4\_1}\right) \tag{3}$$

$$\boldsymbol{F}_r^{5\_1}=V_{\mathrm{encoder}}\left(\boldsymbol{I}_{\mathrm{C}},\mathrm{ReL}\,U_{5\_1}\right) \tag{4}$$

式中：$\boldsymbol{F}_r^{3\_1}$、$\boldsymbol{F}_r^{4\_1}$、$\boldsymbol{F}_r^{5_1}$表示通过不同层提取的图像特征，$V_{\mathrm{encodr}}(\cdot)$表示图像编码器提取图像特征的过程。图像编码器是通过不同卷积层完成图像特征提取，而 ReLU_{3_1}、ReLU_{4_1} 以及 ReLU_{5_1} 是 VGG 网络中的特征提取层，其中 ReLU_{3_1} 用于提取低级视觉特征，ReLU_{4_1} 用于提取中级视觉特征，ReLU_{5_1} 用于提取高级视觉特征，因而不同层能够得到不同层次的特征图并为后续风格迁移提供支撑。

经过多层次特征提取后，多层图像特征$\boldsymbol{F}_r^{3_1}$、$\boldsymbol{F}_r^{4_1}$、$\boldsymbol{F}_r^{5_1}$和风格特征$\boldsymbol{f}_{\mathrm{S}}$一同进入跨模态注意力模块，通过图文交叉融合生成风格引导信号$\boldsymbol{f}_{\mathrm{att}}$并将风格信息注入到图像中，从而实现精细的风格迁移。风格引导信号生成的过程表示为

$$\boldsymbol{f}_{\mathrm{att}}=A\left(\boldsymbol{F}_r^{3\_1},\boldsymbol{F}_r^{4\_1},\boldsymbol{F}_r^{5\_1},\boldsymbol{f}_{\mathrm{S}}\right) \tag{5}$$

式中：$A(\cdot)$表示跨模态注意力机制，通过计算图像特征和风格文本特征之间的关系，生成风格引导信号。

风格引导信号$\boldsymbol{f}_{\mathrm{att}}$被送入1×1卷积层并与图像特征$F_r^{3_1}$进行融合，1×1卷积操作再通过逐元素加和得到同时包含图像结构信息与语义风格信息的特征图F_{CSC}，并为后续扩散优化提供输入：

$$\boldsymbol{F}_{\mathrm{CSC}}=\boldsymbol{F}_{\mathrm{C}}+\boldsymbol{W}_{\mathrm{CS}}\boldsymbol{f}_{\mathrm{att}} \tag{6}$$

式中："+"表示逐元素加和，$\boldsymbol{W}_{\mathrm{CS}}$表示学习得到的权重矩阵，用于调节风格特征的影响。

经过1×1卷积的风格化特征$\boldsymbol{F}_{\mathrm{CSC}}$被传入潜空间扩散模型，潜空间扩散模型通过逐步去噪优化图像特征$\boldsymbol{F}_{\mathrm{out}}$，最终生成风格化图像，其过程可以表示为

$$\boldsymbol{F}_{\mathrm{out}}=L_l(\boldsymbol{F}_{\mathrm{CSC}}) \tag{7}$$

潜在空间扩散模型输出的风格化图像特征$\boldsymbol{F}_{\mathrm{out}}$被传入解码器模块，解码器将这些潜特征转换为最终的风格化图像$\boldsymbol{I}_{\mathrm{CS}}$：

$$\boldsymbol{I}_{\mathrm{CS}}=D(\boldsymbol{F}_{\mathrm{out}}) \tag{8}$$

2.2　自适应模块

为实现更细致的图文融合调控，本文提出自适应注意力归一化模块，该模块通过构建图像与文本之间的跨模态注意力交互生成风格引导信号，并对内容特征进行调整，从而完成风格迁移，如图 2 所示。

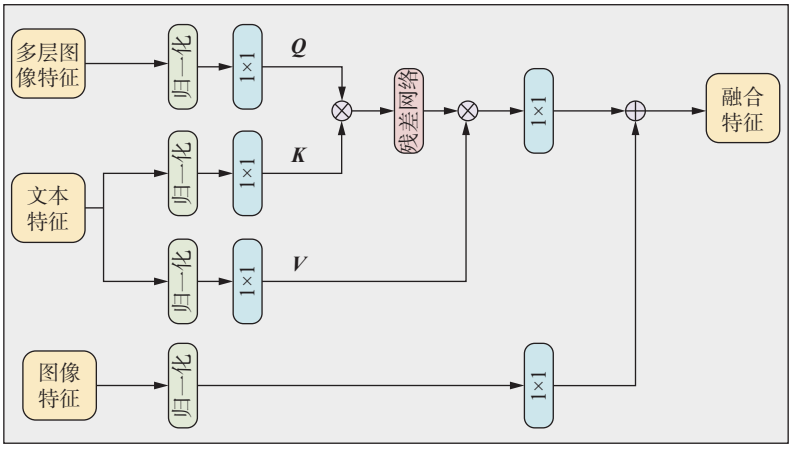

图 2　自适应注意力模块

Fig. 2　Adaptive attention module

首先，内容图像$\boldsymbol{I}_{\mathrm{C}}$和风格文本描述$\boldsymbol{T}_{\mathrm{S}}$分别通过图像编码器与文本编码器来完成特征的提取，得到对应的内容特征$\boldsymbol{F}_{\mathrm{C}}^{X}$和风格特征$\boldsymbol{f}_{\mathrm{S}}$。这些特征再经过归一化的操作，能让它们在数值范围上保持一致。

$$\overline{\boldsymbol{F}}_{\mathrm{C}}^{X}=\mathrm{Norm}\left(\boldsymbol{F}_{\mathrm{C}}^{X}\right),\overline{\boldsymbol{f}}_{\mathrm{S}}=\mathrm{Norm}\left(\boldsymbol{f}_{\mathrm{S}}\right) \tag{9}$$

式中：$\boldsymbol{F}_{\mathrm{C}}^{X}$表示来自图像编码器的内容图像特征；$\boldsymbol{f}_{\mathrm{S}}$表示风格文本描述的特征，可以通过文本编码器 (CLIP) 提取。

图像特征与风格文本特征会先通过卷积操作进行维度调整，使二者的通道数保持一致，从而为后续注意力计算提供条件。

特征映射$\boldsymbol{F}_{\mathrm{r}}^{X}$被作为查询向量$\boldsymbol{Q}$，风格特征$\boldsymbol{f}_{\mathrm{S}}$被映射为对应的键$\boldsymbol{K}$与值$\boldsymbol{V}$。注意力权重通过$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$计算并作用于$\boldsymbol{V}$，生成跨模态风格引导表示$\boldsymbol{f}_{\mathrm{attn}}$：

$$\boldsymbol{Q}=\mathrm{Conv}_1\left(\boldsymbol{F}_{\mathrm{r}}^{X}\right) \tag{10}$$

$$\boldsymbol{K}=\mathrm{Conv}_2\left(\boldsymbol{f}_{\mathrm{S}}\right) \tag{11}$$

$$\boldsymbol{V}=\mathrm{Conv}_3(\boldsymbol{f}_{\mathrm{S}}) \tag{12}$$

$$\boldsymbol{A}_{\mathrm{ttn}}=\mathrm{Softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\right),\boldsymbol{f}_{\mathrm{attn}}=\boldsymbol{A}_{\mathrm{ttn}}\times\boldsymbol{V} \tag{13}$$

风格引导信号与原始图像特征通过 1×1 卷积进行加权融合, 并生成风格化特征图$\boldsymbol{F}_{\mathrm{CSC}}^{X}$, 从而为后续扩散建模提供输入。

$$\boldsymbol{F}_{\mathrm{CSC}}^{X}=\boldsymbol{F}_{\mathrm{r}}^{X}+\boldsymbol{W}_{\mathrm{CS}}\boldsymbol{f}_{\mathrm{attn}} \tag{14}$$

式中: $\boldsymbol{W}_{\mathrm{CS}}$表示学习到的权重矩阵, 该矩阵用于调节风格的特征影响。

该模块实现了图像语义结构与风格语言信号的融合控制, 并为后续扩散阶段提供了兼具结构信息与风格协同性的表示基础。

2.3　潜在空间扩散模型

为进一步提升高分辨率图像风格迁移的质量与效率, 本研究引入潜在扩散模型, 在潜空间里对图像特征演化过程开展建模, 如图 3 所示, 在低维潜在空间当中进行扩散进程, 避免于像素空间进行高维计算, 大幅增进了计算效率, 有效保留图像细节以及风格一致性。

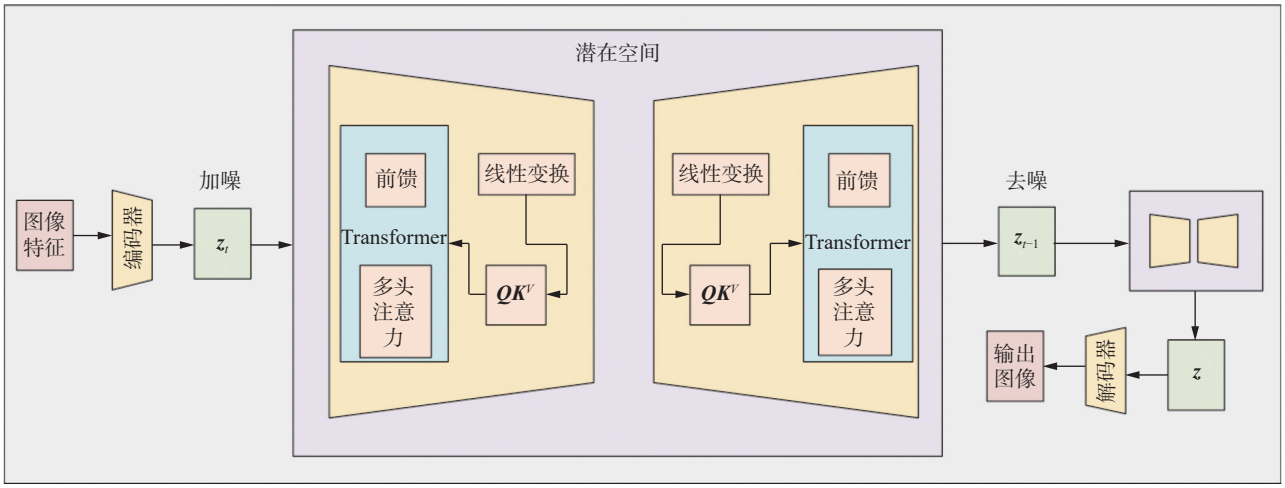

图 3　潜在空间模块

Fig. 3　Latent space module

潜在空间扩散模型以通过前向加噪和反向去噪两阶段来实现, 将图像潜空间演化路径建模来逐步生成风格化图像, 并且借助 Transformer 让跨模态图文特征完成精细交互, 从而保证风格一致性。

图像$\boldsymbol{I}_{\mathrm{C}}$和风格文本描述$\boldsymbol{T}_{\mathrm{S}}$分别通过图像编码器和文本编码器, 得到对应的潜在特征$\boldsymbol{F}_{\mathrm{C}}^{X}$和$\boldsymbol{f}_{\mathrm{S}}$, 这些潜在表示将被送入扩散过程进行处理。潜在空间扩散过程包含加噪过程和去噪过程。

在扩散模型的加噪过程中, 图像的潜在特征$\boldsymbol{z}_0$逐步添加噪声, 直到变成纯噪声$\boldsymbol{z}_T$。该过程建模为

$$q(\boldsymbol{z}_t\mid\boldsymbol{z}_{t-1})=N\left(\boldsymbol{z}_t;\mu_\theta(\boldsymbol{z}_{t-1},t),\sigma_t^2\right) \tag{15}$$

式中: $\boldsymbol{z}_t$表示第t步的潜在表示, $\mu_\theta(\boldsymbol{z}_{t-1},t)$表示由去噪网络预测的均值, σ_t^2表示噪声的方差。

在去噪过程中, 本文凭借去噪网络从噪声潜在表示$\boldsymbol{z}_T$中得出图像的潜在特征$\boldsymbol{z}_0$, 去噪过程公式为

$$p_\theta(\boldsymbol{z}_{t-1}\mid\boldsymbol{z}_t)=N\left(\boldsymbol{z}_{t-1};\mu_\theta(\boldsymbol{z}_t,t-1),\sigma_{t-1}^2\right) \tag{16}$$

通过去噪网络的逐步去噪, 图像的潜在表示会逐步恢复并最终生成风格化图像。为提升跨模态风格一致性, 在扩散过程中引入 Transformer 结构并建模图像内容与语言风格之间的语义对齐关系。图像特征$\boldsymbol{F}_{\mathrm{C}}^{X}$通过线性层计算得到查询向量$\boldsymbol{Q}$, 风格文本特征$\boldsymbol{f}_{\mathrm{S}}$通过线性层计算得到键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$:

$$\boldsymbol{Q}=\mathrm{Linear}\left(\boldsymbol{F}_{\mathrm{C}}^{X}\right) \tag{17}$$

$$\boldsymbol{K}=\mathrm{Linear}(\boldsymbol{f}_{\mathrm{S}}) \tag{18}$$

$$\boldsymbol{V}=\mathrm{Linear}(\boldsymbol{f}_{\mathrm{S}}) \tag{19}$$

通过多头自注意力机制, Transformer 能够捕捉图像和文本特征之间的关系, 生成风格引导信号$\boldsymbol{f}_{\mathrm{att}}$。具体的自注意力计算公式为

$$\mathrm{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{20}$$

式中: $\sqrt{d}$是对$\boldsymbol{Q}$、$\boldsymbol{K}$的维度缩放因子, Softmax 函数计算$\boldsymbol{Q}$和$\boldsymbol{K}$之间的相似度, 从而对$\boldsymbol{V}$加权, 得到最终的风格引导信号。

在 Transformer 中, 经过去噪处理的图像特征通过前馈网络来处理, 用以进一步加工和加强特征的表达:

$$\mathrm{FFN}(x)=\max(0,W_1x+b_1)W_2+b_2 \tag{21}$$

经 Transformer 模块生成的风格引导信号$\boldsymbol{f}_{\mathrm{att}}$被注入到图像潜在特征里面, 风格引导信号和图像特征用1×1卷积进行融合, 得到风格化图像特征$\boldsymbol{F}_{\mathrm{CSC}}^{X}$:

$$\boldsymbol{F}_{\mathrm{CSC}}^{X}=\boldsymbol{F}_{\mathrm{r}}^{X}+\boldsymbol{W}_{\mathrm{CS}}\boldsymbol{f}_{\mathrm{attn}} \tag{22}$$

式中: $\boldsymbol{W}_{\mathrm{CS}}$表示学习到的权重矩阵, 用于调节风格特征对图像特征的影响。

风格化图像特征$\boldsymbol{F}_{\mathrm{CSC}}^{X}$输入去噪模块并通过逐步去噪恢复图像的细节。去噪过程逐步去除噪

声, 最终生成风格化图像的潜在表示$\boldsymbol{F}_{\text{out}}$。

去噪过程的优化公式为

$$\boldsymbol{F}_{\text{out}} = L_{\text{LDM}}\left(\boldsymbol{F}_{\text{CSC}}^{X}\right) \tag{23}$$

经过去噪的潜在特征$\boldsymbol{F}_{\text{out}}$被送往解码器并生成风格化图像$\boldsymbol{I}_{\text{CS}}$。

在融合图像结构和风格语义所形成的潜空间中实现图像渐进式生成, 并与前端语义引导以及区域控制模块共同协作, 有力提高最终生成图像的风格表现力和结构保真程度。

2.4　损失函数

为有效训练本文的多阶段风格迁移框架, 本文设计了多组损失函数, 以优化图像的风格一致性、内容保真度与高分辨率生成质量。损失函数主要包含内容损失、风格损失、对抗性损失, 以及潜在空间扩散模型去噪过程对应的去噪损失。

内容损失用于确保生成图像的内容特征与原始内容图像保持一致。在本框架中, 通过图像编码器提取图像特征表示$\boldsymbol{F}_{\text{C}}^{X}$, 计算生成图像与原始内容图像的特征差异, 以最小化图像内容变化。具体而言, 内容损失通过对比内容图像与生成图像在相同层次的特征表示进行计算。

$\boldsymbol{F}_{\text{C}}^{X}$表示原始内容图像的特征, $\boldsymbol{F}_{\text{gen}}^{X}$表示生成图像的特征, 内容损失函数$L_{\text{C}}$表示为

$$L_{\text{C}} = \frac{1}{N}\sum_{i}\left\|\boldsymbol{F}_{\text{C}}^{X} - \boldsymbol{F}_{\text{gen}}^{X}\right\|_{2}^{2} \tag{24}$$

式中: N表示图像特征的维度, $\|\cdot\|_2$表示欧几里得距离。

风格损失用于保障生成图像风格与目标风格描述匹配, 通过文本编码器提取风格文本特征$\boldsymbol{f}_{\text{S}}$, 对比生成图像与目标风格图像的 Gram 矩阵计算损失, 量化风格差异。

$\boldsymbol{G}_{\text{C}}$表示内容图像的 Gram 矩阵, $\boldsymbol{G}_{\text{gen}}$表示生成图像的 Gram 矩阵, 风格损失函数$L_{\text{S}}$表示为

$$L_{\text{S}} = \frac{1}{N}\sum_{i}\left\|\boldsymbol{G}_{\text{gen}} - \boldsymbol{G}_{\text{C}}\right\|_{2}^{2} \tag{25}$$

式中N表示 Gram 矩阵的维度。通过最小化风格损失, 本文确保生成的图像不仅保持原有的内容, 还能准确反映目标风格的语义信息。

为进一步提升生成图像的真实感与细节质量, 本文引入对抗性损失。借助对抗训练, 可促使生成模型学习更细腻的风格细节; 损失通过生成对抗网络 (GANs) 实现, 由生成器生成风格化图像, 判别器负责判定生成图像的真实性。

生成器通过最小化对抗性损失来提高图像的生成质量, 对抗性损失L_{a}可表示为

$$L_{\text{a}} = \text{E}_{I_{\text{gen}}}\left[\log\left(1 - D\left(\boldsymbol{I}_{\text{gen}}\right)\right)\right] \tag{26}$$

式中: $D\left(\boldsymbol{I}_{\text{gen}}\right)$为判别器输出的对抗判定值, 用于表征生成图像$\boldsymbol{I}_{\text{gen}}$是否为真实图像。生成器的核心目标是让判别器将生成图像判定为 “真实”, 进而最小化对抗性损失。

在潜在空间扩散模型中, 去噪损失是不可或缺的组成部分, 其直接影响图像细节恢复效果与风格一致性。本文通过去噪网络 (UNet) 对潜在空间中的噪声进行逐步去噪, 确保图像在生成过程中既能维持风格一致性, 又能实现细节的精细化呈现。

去噪损失通过计算生成图像潜在特征和目标风格化图像潜在特征之间的差异来衡量。$\boldsymbol{z}_0$为去噪后的图像潜在特征, $\boldsymbol{z}_{\text{gen}}$为生成图像的潜在特征, 去噪损失$L_{\text{d}}$表示为

$$L_{\text{d}} = \frac{1}{N}\sum_{i}\left\|\boldsymbol{z}_0 - \boldsymbol{z}_{\text{gen}}\right\|_{2}^{2} \tag{27}$$

通过最小化去噪损失生成图像能够符合目标风格, 并且可以恢复较高质量的图像细节。最终, 损失函数L_{t}是各子损失函数加权后的结果, 结合了内容损失、风格损失、对抗性损失以及去噪损失, 可以表示为

$$L_{\text{t}} = \lambda_{\text{C}}L_{\text{C}} + \lambda_{\text{S}}L_{\text{S}} + \lambda_{\text{a}}L_{\text{a}} + \lambda_{\text{d}}L_{\text{d}} \tag{28}$$

式中: λ_{C}、λ_{S}、λ_{a}、λ_{u}表示权重系数, 用于平衡各个损失函数在训练过程中的影响。

为了提高图像风格迁移任务里解决语义引导不确切、区域风格控制粗糙以及高分辨率图像细节丢失等问题的能力, CAST-Diff 框架在 4 个方面开展了结构设计并进行机制创新: 1) 三阶段协同结构设计, 创建了由“跨模态语义引导—区域风格调节—潜空间扩散重构”所构成的风格迁移流程, 把语义一致性、局部风格可控性和细节保真度组合成统一优化目标; 2) 区域感知相关的风格调节机制, 提出一种自适应归一化风格模块, 能根据图像区域特征动态管理风格强度, 大幅提高风格过渡的自然性和局部连贯性; 3) 后置式扩散生成路径设定, 与传统扩散方法在做法上不同, CAST-Diff 把扩散建模用于风格融合之后的图像后期细化阶段, 并于潜空间加以实施, 以此在保障细节重建质量的基础上降低训练资源消耗; 4) 跨模态控制嵌入实现生成链条全程覆盖, 本文把 CLIP 编码的风格引导信号深度投入扩散去噪过程, 使文本语义控制贯穿整个生成的路径, 突破了传统风格迁移只靠前端对齐的局限。

通过引入跨模态语义引导、区域风格调节、潜扩散细化机制来做解耦式协同设计, 从理论层面看, CAST-Diff 可保证任务驱动的语义对齐、迁移强度可控以及细节重建效率这三者的动态平衡, 在结构机制上有力提升模型的泛化能力及稳定性。

3　实验

3.1　实验设置

本文在 COCO 和 Flickr30k 两个数据集进行实验, COCO 数据集有 123 287 张图像, 包含 80 种物体类别, 每张图像都拥有文本描述, 因此经常用于图像生成以及风格迁移相关任务; Flickr30k 数据集包含 31 000 张图像, 每张图像都有对应的文本描述。开展实验时, 所有图像都被改成 256×256 像素, 并且采用随机裁剪、翻转以及颜色抖动等方式来做数据增强预处理。

所有实验均在 NVIDIA GeForce RTX 4090 GPU 上使用 PyTorch1.10 框架进行。训练过程中, 使用 Adam 优化器, 学习率为 0.000 1, 批量大小为 16, 训练周期为 50 000 次。

3.2　定性评价

从图 4 可以看出, CAST-Diff 在风格迁移任务中能够有效地将内容图像与目标风格进行融合。生成的图像在风格表现上保持了一定的一致性, 且细节得到了较好的保留。尤其在风景图像的处理上, CAST-Diff 能够较好地处理图像的复杂背景, 保持风格和内容的平衡。人物图像的风格迁移效果较为自然, 面部细节和表情得到了一定的保留。

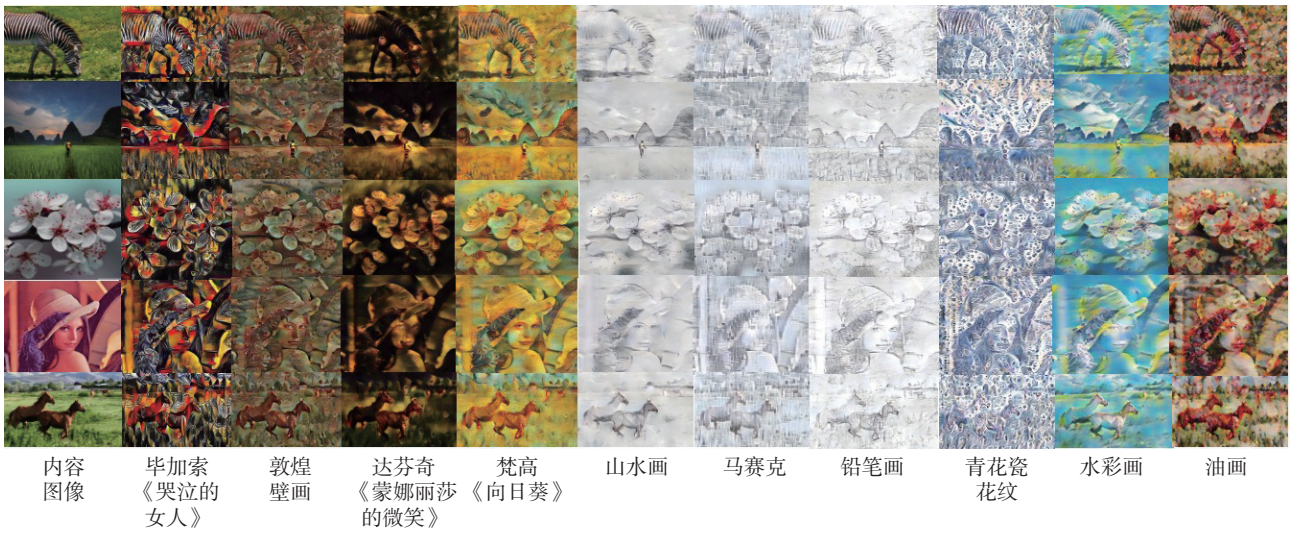

图 4　CAST-Diff 风格迁移效果

Fig. 4　CAST-Diff style transfer results

尽管风格迁移的效果较为自然, 但生成的图像仍存在风格过渡不够平滑以及局部细节模糊的情况。因此 CAST-Diff 在风格一致性和细节保留方面具有优势, 但在处理复杂场景时仍有进一步优化的空间。

3.2.1　与传统风格迁移方法比较

为了评估本文的方法, 本文与几种传统的风格迁移方法进行了对比, 包括 ArtFlow[26]、IEContraAST[12]、AesUST[3]、SD-Net[4]、SANet[29] 等。图 5 给出了不同方法在风格迁移任务中的生成结果。

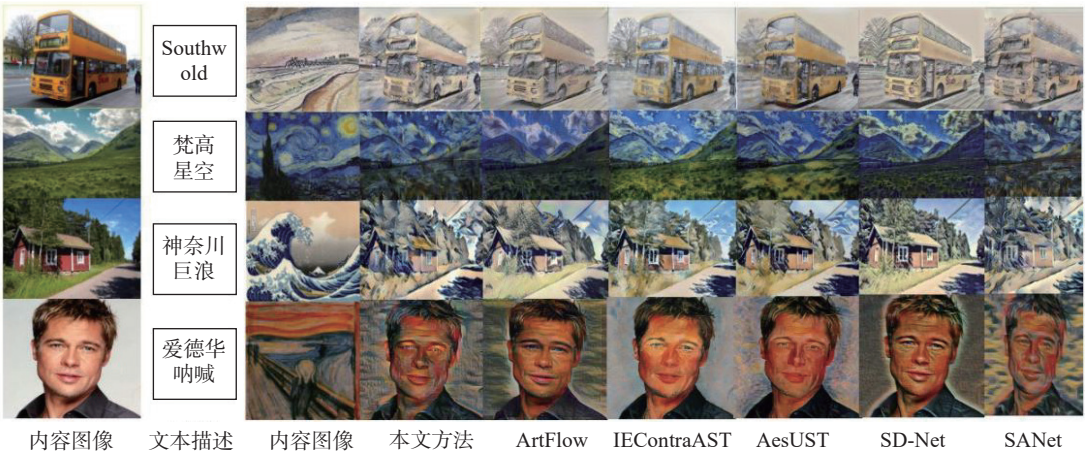

图 5　与传统风格迁移方法对比

Fig. 5　Compared with traditional style transfer methods

在第一个示例中，ArtFlow 在风格迁移过程中虽能生成较为平滑的图像，但处理复杂背景与细节时往往缺乏足够的风格层次，导致风格迁移效果略显平淡。CAST-Diff 生成的图像不仅能更好地保留原始图像细节，在风格过渡与内容表达上也更显自然细腻。尤其在人物面部特征保留上，CAST-Diff 可有效避免面部特征模糊，在保证风格一致性的同时确保面部细节清晰可辨。

在处理背景复杂的风景图像时，IEContraAST[17] 也能呈现出较为平滑的风格迁移效果，但当背景中包含更多复杂元素时，其风格迁移的自然性会受到一定影响。CAST-Diff 可更好地实现风格与内容的融合，有效避免过度迁移及风格不一致问题，生成图像的层次感更鲜明、风格一致性更优。

CAST-Diff 相较于传统风格迁移方法，在风格一致性、细节保真度以及视觉呈现效果上展现出显著优势。在处理复杂背景及细节丰富的图像时，本文所提方法能够更好地保留图像的内容结构以及风格一致性，输出更自然精细的风格迁移结果。

3.2.2　与文本引导生成风格迁移方法比较

为了进一步评估本文方法的效果，本文与当前流行的文本引导生成风格迁移方法进行对比，包括 T2I Adapter[20]、ControlNet[25]、InST[30]、IEST[17] 和 StyleStudio[31]。图 6 给出了不同方法在相同内容图像和风格描述下的生成结果。

图 6　与文本引导生成风格迁移方法对比

Fig. 6　Comparison with text-guided generative style transfer methods

T2I Adapter 与 ControlNet 虽可根据文本描述实现基础风格迁移，但风格过渡生硬且内容保真度与细节清晰度较差。CAST-Diff 在处理文本引导的风格迁移时，能够在风格和内容的融合上取得显著的平衡，生成的图像风格更加自然细腻，尤其在复杂背景图像中能够更加稳定地保留原图的细节。

此外，InST 与 IEST 虽能生成较为合理的风格迁移结果，但在处理复杂图像内容时，仍存在风格与内容耦合度不足的问题。CAST-Diff 可实现风格与内容的无缝过渡，保证生成图像的自然性与一致性。

与现有文本引导的风格迁移方法相比，CAST-Diff 在风格一致性、细节保真度与风格过渡自然性上均表现出显著优势。在复杂背景与细节区域的处理中本文方法可更好地实现风格与内容的深度融合，生成视觉效果更自然、细腻的图像。相较于 T2I Adapter、ControlNet 等方法，CAST-Diff 能更大程度保留原图细节，保证风格迁移的一致性与自然度。

3.2.3　与代表性方法的结构对比分析

为进一步突出 CAST-Diff 在结构设计以及核心机制的独特表现，本文和当前流行的风格迁移方法进行了对比分析。当前方法倾向于图像结构控制以及语义引导建模，多阶段协同机制和区域级风格调节能力仍有显著的不足，与 CAST-Diff 的设计理念有显著的差别。

StyleGAN-Diffusion 把全局风格控制作为重点，缺乏区域细粒度的调节；ControlNet 可以凭借视觉提示把控图像的结构，语义风格自适应调节的限度较大；DALLE-style 依靠预训练使文本驱动能够实现，没有动态控制能力；WAN 实施多阶段特征融合，但没有引入潜扩散机制，细节保真效果不好。

CAST-Diff 以图文融合、多尺度风格调节和潜空间精细生成的方式，弥补了当前方法在区域

风格控制以及语义一致性方面的空缺。

3.3　定量比较

3.3.1　用户评价

为进一步验证 CAST-Diff 方法在主观视觉质量方面的优势, 本文设计了用户偏好调查。参与者在不知道方法名称的情况下, 依据生成图像的风格一致性、细节保真度以及整体视觉效果进行选择, 并从中选出自己更偏好的图像。

本文共邀请 30 名参与者, 其中包括计算机视觉领域研究人员以及普通用户。每位参与者都会在 COCO 与 Flickr30k 测试集中随机抽取 50 组样本进行评价, 每组样本包含 11 种不同方法的生成结果, 同时展示顺序被随机打乱, 避免潜在偏倚。

最终统计了各方法在所有样本中被选为首选的比例, 结果如图 7 所示。

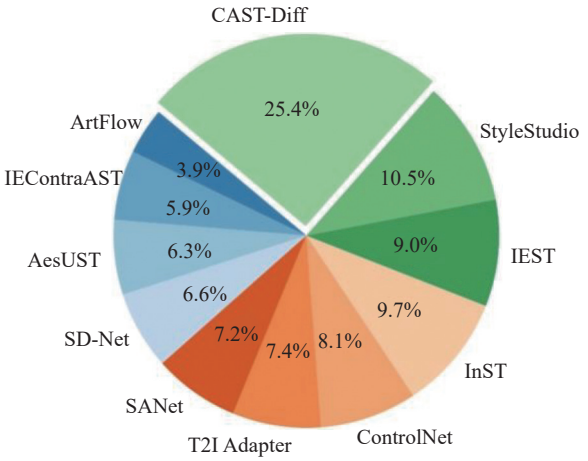

图 7　用户评价结果

Fig. 7　User evaluation results

从图 7 可以看出, CAST-Diff 以 25.4% 的选择率高于其他方法。相比之下, 表现较好的 StyleStudio、InST 和 IEST 分别获得了 10.5%、9% 和 9.7% 的选择率。CAST-Diff 在风格一致性、细节保真度以及整体观感上的优势, 使其在用户主观评价中仍然获得了最高比例的偏好。

用户反馈显示, CAST-Diff 生成图像在风格与内容的细节丰富度及整体视觉流畅性上表现更优, 尤其在处理复杂场景时, 能更好保留清晰结构并实现柔和风格迁移。

3.3.2　性能比较

为了系统评估所提出的 CAST-Diff 方法在文本引导风格迁移任务中的性能, 本文在 COCO 和 Flickr30k 数据集上开展全面测试。平均结果如表 1 所示。评估采用了 FID(fréchet inception distance)、LPIPS(learned perceptual image patch similarity)、PSNR(peak signal-to-noise ratio) 和 SSIM (structural similarity index measure)4 项指标, 分别衡量生成图像的感知质量、风格一致性与内容保真性。较低的 FID 和 LPIPS 表明生成图像在感知质量以及风格表达方面更具优势, 较高的 PSNR 和 SSIM 表明内容结构保留更好。

表 1　与各类风格迁移方法的质量评价表

Table 1　Quality assessment table for various style transfer methods

方法	FID(↓)	LPIPS(↓)	PSNR(↑)	SSIM(↑)
ArtFlow[26]	32.7	0.279	25.5	0.811
IEContraAST[17]	30.8	0.268	25.9	0.817
AesUST[3]	29.9	0.260	26.2	0.823
SD-Net[4]	29.1	0.252	26.6	0.831
SANet[30]	28.4	0.248	27.0	0.836
T2I Adapter[12]	27.6	0.241	27.4	0.842
ControlNet[25]	26.7	0.236	27.6	0.844
InST[31]	26.1	0.237	27.9	0.848
IEST[20]	25.7	0.234	28.1	0.853
StyleStudio[3]	25.1	0.229	28.3	0.851
本文方法	24.4	0.231	28.7	0.852

从表 1 可见, CAST-Diff 方法在 FID 与 PSNR 两项指标上取得了最优结果, 且明显优于其他基准方法。CAST-Diff 在感知质量指标 FID 上比 StyleStudio 进一步降低 0.7, 在内容保真性指标 PSNR 上提升 0.4, 说明该方法在风格迁移过程中能够较好兼顾风格一致性以及细节保留。

ControlNet 与 IEST 虽然分别在 LPIPS 和 SSIM 指标上略占优势, 但 CAST-Diff 在各项指标上的整体表现更稳定。尤其在高分辨率图像以及复杂背景条件下, CAST-Diff 生成的图像在细节还原、风格融合以及整体视觉质量方面更自然协调。

CAST-Diff 在 COCO、Flickr30k 等不同数据集上都保持了稳定并且一致的领先表现, 说明本文提出的跨模态引导、自适应风格调节以及扩散优化机制在文本引导图像风格迁移任务中具有较好的有效性与鲁棒性。

3.4　消融实验

为验证 CAST-Diff 各模块对生成效果的实际贡献, 本文设计了消融实验, 逐步把模型中的不同模块移除, 仔细观察生成图像所产生的变化。为进一步验证模型在不同结构配置下的稳定性与鲁棒性, 本文针对相关模块与损失函数开展了以下消融实验。通过对比每种消融设置下生成的图像, 本文评估了每个模块对最终生成图像质量的影响, 这些消融实验也帮助本文清晰理解不同模块对模型性能的具体作用, 充分展现了每个模块在提升图像生成质量的贡献。

3.4.1　损失函数消融实验

图 8(a) 给出了去除不同损失函数之后的消融实验结果。图 8(b) 去除内容损失会使图像结构模糊、细节出现缺失，还会降低图像的真实感；图 8(c) 缺少了风格损失又会使图像的风格表达不够自然，呈现出简单的叠加效果，很难实现风格与内容之间的有效融合；图 8(d) 缺少潜在空间损失则会削弱图像的细节表现与结构层次，影响到风格和内容之间的协调程度；相比之下，完整保留了 3 种损失项的最终生成图像，在风格一致性、结构保真度和整体视觉质量方面的表现都是最好的，也验证了各个损失项在模型优化过程中起到的关键作用。

3.4.2　模块组件消融实验

图 9 给出了 CAST-Diff 各模块的消融实验效果。去除跨模态注意力模块 (图 9(a)) 后，风格特征模糊、过渡不自然。去除自适应风格调节模块 (图 9(b)) 会导致局部风格控制力减弱，图像细节丢失，表现出风格迁移强度不均的现象。去除潜在空间模块 (图 9(c)) 则使图像整体质感下降，细节层次感削弱。相比之下，保留全部模块的最终生成图像 (图 9(d)) 在风格一致性、细节保真度与融合自然性方面表现最优，验证了各模块协同作用的重要性。

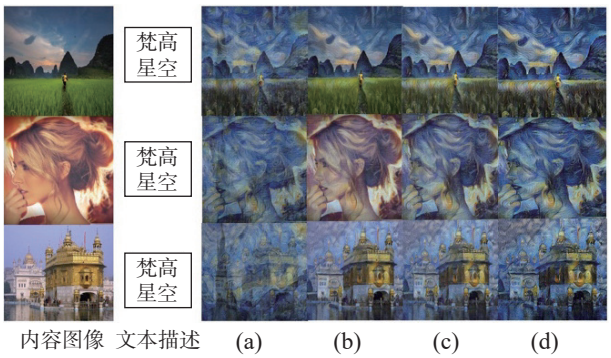

图 8　损失消融实验

Fig. 8　Loss ablation experiment

图 9　各个模块组件消融实验

Fig. 9　Ablation experiment of each module component

3.5　计算效率分析

为评估模型的计算资源开销，本文在标准硬件平台下从推理时间、GPU 显存占用和模型吞吐量 3 个方面进行测试。为了全面评估所提出的 CAST-Diff 方法在风格迁移任务中的计算效率，本文在 NVIDIA GeForce RTX 4090 GPU 上与几种基准方法进行了对比。本文从推理时间、GPU 内存占用和吞吐量 3 个角度进行评估，如表 2 所示。所有方法均使用相同的图像分辨率 (256×256) 进行生成，确保了实验的公平性和可比性。

在推理时间方面，CAST-Diff 每张图像的生成时间有 2.7 s，和其他相关方法比起来表现处于中等水平，但与 StyleStudio(3.5 s) 相比，它的推理时间有了明显减少，这也表明本文所提出的模型在保证好图像质量的同时，还能在合理的时间内生成图像；虽然 ControlNet(3.2 s) 和 AesUST(3.0 s) 的推理时间略长，但这两种方法也呈现出了较高的生成质量，这一点也为模型在计算效率与生成质量之间的权衡提供了相应的参考。

在 GPU 内存占用方面，CAST-Diff 显存占用为 1 900 MB，处于基准方法中上游；相较于 ArtFlow (1 200 MB) 和 IEContraAST(1 350 MB)，其显存占用更高，这是由于多阶段风格迁移优化与跨模态注意力机制增加了内存消耗，但该显存占用仍在可接受范围，且能有效保持生成图像的细节与风格一致性。

在吞吐量方面，CAST-Diff 每秒可生成 0.22 张图像，相较于其他方法略低，但这并未影响其图像质量表现；生成高质量图像通常需要更多推理时间与计算资源，因此适度降低吞吐量是可

接受的, 尤其在对图像质量要求较高的应用场景中。其吞吐量与 StyleStudio(0.22 张/s) 几乎持平, 两者生成速度差距较小。

表 2　效率分析表

Table 2　Efficiency analysis table

方法	推理时间/s	GPU内存占用/MB	吞吐量/(张/s)
ArtFlow[26]	2.3	1 200	0.31
IEContraAST[17]	2.7	1 350	0.29
AesUST[3]	3.0	1 500	0.26
SD-Net[4]	2.8	1 450	0.28
SANet[30]	2.6	1 400	0.30
T2I Adapter[12]	2.9	1 450	0.27
ControlNet[25]	3.2	1 600	0.24
InST[31]	2.8	1 500	0.29
IEST[20]	2.6	1 450	0.30
StyleStudio[3]	3.5	1 700	0.22
本文方法	2.7	1 900	0.22

CAST-Diff 在推理时间与吞吐量方面虽然略弱于其他方法, 但该方法在风格一致性以及图像细节保留方面表现更好, 因而在高质量风格迁移任务中仍具有较强竞争力。其内存占用较高且推理时间适中, 并且更适合对风格迁移质量要求较高的应用场景。

4　结束语

本文提出一种新的多阶段风格迁移框架 CAST-Diff, 该框架融合了跨模态注意力机制、自适应风格控制以及扩散建模, 并通过对图像内容与风格信息融合过程的细致调控, 使 CAST-Diff 在多个基准数据集上都表现出较高的生成质量、更强的风格一致性以及更好的细节保真度。

在定性评价部分, 本文将 CAST-Diff 与传统风格迁移方法以及文本引导生成方法进行对比, 结果表明该模型在处理复杂背景并且细节丰富的图像时, 能够更好保留图像结构并维持风格一致性, 从而生成更自然、更细致的风格迁移结果。在多模态输入条件下, CAST-Diff 还能更充分挖掘图像与文本之间的潜在关联, 使风格引导与表达能力进一步增强。

在定量评估部分, 本文进一步验证了 CAST-Diff 在风格迁移任务中的良好效果, 该模型在指标上都表现较好, 并且整体结果优于现有主流方法。虽然 CAST-Diff 在计算效率方面存在一定代价, 但该模型与其他高质量生成方法相比, 其推理时间以及内存占用仍处于可接受范围, 因此更适合对生成质量要求较高的风格迁移应用。

在消融实验部分, 本文验证了各模块在 CAST-Diff 中的关键作用。任意一个模块被去除后都会导致生成图像质量下降, 因此跨模态注意力机制以及自适应风格控制模块对于保证风格一致性与细节保真度是十分重要的。

CAST-Diff 在维持图像质量的基础上能够有效提升风格迁移的稳定性, 并且适用于图像生成、艺术创作以及个性化设计等实际场景。未来工作可继续围绕计算效率展开优化, 提升推理速度以及吞吐量, 从而适配更多样化且更具实时性的应用需求。

参考文献:

[1] YOSHIOKA D, YASUDA Y, TODA T. Nonparallel spoken-text-style transfer for linguistic expression control in speech generation[J]. IEEE transactions on audio, speech and language processing, 2025, 33: 333−346.

[2] LI Xiangtian, CAO Han, ZHANG Zhaoyang, et al. Artistic neural style transfer algorithms with activation smoothing[C]//Proceedings of the 2025 2nd International Conference on Informatics Education and Computer Technology Applications. New York: ACM, 2025: 1−6.

[3] WANG Zhizhong, ZHANG Zhanjie, ZHAO Lei, et al. AesUST: towards aesthetic-enhanced universal style transfer[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1095−1106.

[4] WANG Quan, LI Sheng, WANG Zichi, et al. Multi-source style transfer via style disentanglement network[J]. IEEE transactions on multimedia, 2024, 26: 1373−1383.

[5] JING Yongcheng, YANG Yezhou, FENG Zunlei, et al. Neural style transfer: a review[J]. IEEE transactions on visualization and computer graphics, 2020, 26(11): 3365−3385.

[6] CHEN Haibo, ZHAO Lei, WANG Zhizhong, et al. Artistic style transfer with internal-external learning and contrastive learning[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 26561−26573.

[7] DENG Yingying, TANG Fan, DONG Weiming, et al. StyTr2: image style transfer with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11316−11326.

[8] DENG Yingying, TANG Fan, DONG Weiming, et al. Arbitrary style transfer via multi-adaptation network[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 2719−2727.

[9] GATYS L A, ECKER A S, BETHGE M. A neural algorithm of artistic style[EB/OL]. (2015−08−26)[2025−01−01]. https://arxiv.org/abs/1508.06576.

[10] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: ICCV, 2017: 1501−1510.

[11] LI Yijun, FANG Chen, YANG Jimei, et al. Universal

style transfer via feature transforms[C]//Advances in Neural Information Processing Systems. Long Beach: NIPS, 2017: 386−396.
[12] 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. Piscataway: IEEE, 2022: 10674−10685.
[13] ZHANG Yuxin, TANG Fan, DONG Weiming, 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.
[14] LIU Songhua, LIN Tianwei, HE Dongliang, et al. AdaAttN: revisit attention mechanism in arbitrary neural style transfer[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 6629−6638.
[15] 贵向泉, 李琪, 李立, 等. 多尺度感知的单文本条件图像风格迁移[J]. 计算机技术与发展, 2025, 35(9): 46−54.
GUI Xiangquan, LI Qi, LI Li, et al. Single text conditional image style transfer for multi-scale perception[J]. Computer technology and development, 2025, 35(9): 46−54.
[16] 董心悦, 傅鹏. 基于改进生成对抗网络的人脸图像风格迁移方法[J]. 鄂州大学学报, 2025, 32(2): 94−97.
DONG Xinyue, FU Peng. Face image style transfer method based on improved generative adversarial network[J]. Journal of Ezhou University, 2025, 32(2): 94−97.
[17] CHEN Haibo, ZHAO Lei , WANG Zhizhong, et al. Artistic style transfer with internal-external learning and contrastive learning[J]. Advances in neural information processing systems, 2021, 34: 26561−26573.
[18] 雷松林, 赵征鹏, 阳秋霞, 等. 基于可解耦扩散模型的零样本风格迁移[J]. 图学学报, 2025, 46(4): 727−738.
LEI Songlin, ZHAO Zhengpeng, YANG Qiuxia, et al. Zero-shot style transfer based on decoupled diffusion models[J]. Journal of graphics, 2025, 46(4): 727−738.
[19] WANG Zhizhong, ZHAO Lei, ZUO Zhiwen, et al. MicroAST: towards super-fast ultra-resolution arbitrary style transfer[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(3): 2742−2750.
[20] MOU Chong, WANG Xintao, XIE Liangbin, et al. T2I-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models[J]. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(5): 4296−4304.
[21] LEI Mingkun, SONG Xue, ZHU Beier, et al. StyleStudio: text-driven style transfer with selective control of style elements[EB/OL]. (2024−12−11)[2025−01−01]. https://arxiv.org/abs/2412.08503.
[22] GAO Xiang, ZHANG Yuqi. SRAGAN: saliency regularized and attended generative adversarial network for Chinese ink-wash painting style transfer[J]. Pattern recognition, 2025, 162: 111344.
[23] PATASHNIK O, COHEN A, SHECHTMAN E. StyleCLIP: text-driven manipulation of StyleGAN imagery[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2021: 2085−2094.
[24] GAL R, PATASHNIK O, COHEN-OR D. StyleGAN-NADA: CLIP-guided domain adaptation of image generators[EB/OL]. (2021−08−03)[2025−01−01]. https://arxiv.org/abs/2108.00946.
[25] ZHANG Lyumin, RAO Anyi, AGRAWALA M. Adding conditional control to text-to-image diffusion models[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 3813−3824.
[26] AN Jie, HUANG Siyu, SONG Yibing, et al. ArtFlow: unbiased image style transfer via reversible neural flows[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 862−871.
[27] YU Zhenyu, WANG Jinnian, CHEN Hanqing, et al. QRS-trs: style transfer-based image-to-image translation for carbon stock estimation in quantitative remote sensing[J]. IEEE access, 2025, 13: 52726−52737.
[28] ZHANG Yuxin, HUANG Nisha, TANG Fan, et al. Inversion-based style transfer with diffusion models[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 10146−10156.
[29] PARK D Y, LEE K H. Arbitrary style transfer with style-attentional networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 5873−5881.
[30] SOHN K, RUIZ N, LEE K, et al. StyleDrop: text-to-image generation in any style[EB/OL]. (2023−06−01)[2025−01−01]. https://arxiv.org/abs/2306.00983.
[31] 王伟, 张静宜, 温玉辉, 等. 基于神经网络的图像风格迁移算法综述[J]. 电子学报, 2025, 53(5): 1692−1712.
WANG Wei, ZHANG Jingyi, WEN Yuhui, et al. Neural network based image style transfer: a survey[J]. Acta electronica sinica, 2025, 53(5): 1692−1712.

作者简介：

仇佳庆，硕士研究生，主要研究方向为多媒体、计算机视觉。E-mail：983576605@qq.com。

窦立云，讲师，博士，主要研究方向为图像处理、图像篡改，先后担任了 ACMMM、CVPR、TMM 和 TKDE 等国际顶会顶刊的审稿人。E-mail：liyun_dou@163.com。

王进，副教授，博士，中国计算机学会高级会员，南通市人工智能学会副理事长，主要研究方向为人工智能、计算机视觉，中国计算机学会 (CCF) 高级会员，南通市人工智能学会副理事长。发表学术论文 40 余篇，申请授权专利 30 余件。E-mail：wj@ntu.edu.cn。