

中图法分类号: TP399 文献标识码: A 文章编号: 1006-8961(2025)09-3141-12

论文引用格式: Zhang Y X, Dong W M and Xu C S. 2025. Arbitrary image style transfer based on swapping attention mechanism. Journal of Image and Graphics, 30(9): 3141-3152 (张宇欣, 董未名, 徐常胜. 2025. 基于交换注意力机制的任意图像风格迁移. 中国图象图形学报, 30(9): 3141-3152) [DOI: 10.11834/jig.240652]

基于交换注意力机制的任意图像风格迁移

张宇欣, 董未名*, 徐常胜

1. 中国科学院自动化研究所, 北京 100190; 2. 中国科学院大学人工智能学院, 北京 100049

摘要: 目的 图像风格迁移旨在将一种艺术风格应用到一幅真实图像上, 同时保持内容的完整性。传统的风格迁移方法存在伪影、风格模糊等问题。文字到图像生成扩散模型难以通过语言精确表达特定作品的创意。现有研究通过微调模型或文本嵌入技术提升风格准确度, 但微调模型引入了额外计算成本, 效率较低。针对这些挑战, 本文提出一种无需额外微调模型的高效风格迁移方法。**方法** 本文方法包含一种新颖的三线并行风格迁移框架, 即图像生成路径、内容引导路径和风格引导路径。利用摄影和艺术图像作为视觉提示, 通过一种新颖的注意力交换机制, 将引导图像信息注入新图像生成路径, 实现灵活、高效的任意图像风格迁移。**结果** 通过定性和定量实验比较, 本文方法能够在推理时间内生成高质量的风格迁移结果, 定性实验表明本文方法适用于人像、风景和静物等多种题材, 水墨、油画和素描等多种风格, 且视觉效果优于对比的先进风格迁移方法; 定量实验表明本文方法提升了风格准确性, 取得了最优结果。**结论** 所提方法在零微调设置下即可对任意输入图像完成风格迁移, 无需针对特定风格或内容重新训练或微调模型。该方法能够同时处理丰富多样的视觉内容, 涵盖人像、建筑、静物、自然风景等常见题材, 以及广泛的艺术风格, 包括但不限于水墨、油画、素描等主流风格。并在上述任意内容与风格的组合中, 均稳定输出高保真、风格一致、语义保持完整的迁移结果。

关键词: 图像生成; 生成式模型; 扩散模型; 风格迁移; 注意力机制

Arbitrary image style transfer based on swapping attention mechanism

Zhang Yuxin, Dong Weiming*, Xu Changsheng

1. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;

2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Objective Generative methods based on diffusion models have garnered widespread attention due to their diverse generation outcomes and high-quality effects. Diffusion models have a distinct advantage over traditional style transfer frameworks in terms of the diversity, quality, and realism of generated images, and are effective in capturing the complex distributions of data. With their robust data modeling capabilities, diffusion models are particularly suitable for multi-modal generation tasks. However, applying diffusion models to text-to-image generation for style transfer tasks still faces several challenges. First, the artistic styles involved in style transfer often possess highly visual and artistic characteristics that are difficult to precisely express through language. Although text-guided stylization techniques can generate artistic images by using natural language and textual prompts, the textual prompts for the target style are often limited to a rough description of materials, artistic genres, artists, or famous artworks. Detailed auxiliary textual input is usually required to

收稿日期: 2024-10-29; 修回日期: 2025-01-19; 预印本日期: 2025-01-26

* 通信作者: 董未名 weiming.dong@ia.ac.cn

基金项目: 北京市自然科学基金项目 (L221013)

Supported by: Beijing Natural Science Foundation, China (L221013)

guide the generation process in reproducing vivid content and style. Even so, the generated results may still fail to fully capture the creativity and conception of a specific painting. Researchers in the field of image generation have explored the use of reference images as visual cues to enhance the quality and diversity of generated images. These methods mainly include fine-tuning diffusion models or text embedding techniques. These techniques are often accompanied by high fine-tuning costs, including the need for a large amount of data and computational resources. **Method** The core of image style transfer lies in two aspects: transferring the artistic style of the image while maintaining the content information of the original image. We designed a novel three-branch parallel generation path, namely, the new image generation path, the content reference path, and the style reference path, with the goal of integrating style and content features into the image generation process. This feature fusion is achieved through an interactive attention module. Specifically, the process begins with three generation paths: One starts from text prompts and initial noise, naturally forming an image that matches the text prompts, while the other two paths are directed toward an artistic image and a content image, respectively. To generate an image that integrates the reference style features, we utilize the hierarchical generation characteristics of the diffusion model, which generates structured information in the deep network of the model and texture, color, and other information in the shallow network, corresponding to the content and style information set in the task. On the basis of this characteristic, in the deep self-attention module of the diffusion model's main network, the key and value features in the original path are exchanged with the corresponding features in the content path. Subsequently, in the shallow self-attention module of the diffusion model's main network, the key and value features in the original path are exchanged with the corresponding features in the style path. In this way, the attention mechanism uses the similarity between the original query features and the reference key features to reconstruct the reference value features in the form of weights. Tone consistency is a key factor in evaluating the effect of style transfer. Although methods based on interactive attention perform well in transferring the brushstrokes and textures of artistic images, they often have difficulty aligning the overall tone of the content image with the style image. The image generation of the diffusion model depends not only on the generation network and text conditions but is also significantly influenced by the initial noise. Previous studies pointed out that the color tone of the generated image is closely related to the sampling of the initial noise. On the basis of this finding, we address the tone consistency issue between the generated image and the style image by adjusting the initial noise. We propose a new method to improve tone consistency by aligning the initial noise of the generated image with the distribution of the style image. **Result** We conducted a comparative analysis with five state-of-the-art traditional style transfer methods, namely, ArtFlow, AdaAttN, StyTr2, CAST, and AesPA-Net. For the test dataset, we utilized a collection of 100 artistic images from WikiArt and 100 real-world images from the Places365 dataset, which were randomly sampled from a dataset exceeding 100 000 images to ensure a fair comparison. With regard to implementation details, we employed stable diffusion extra large (SDXL) with its default hyperparameters throughout all our experiments. Our approach enhanced the style accuracy by 4%, achieving a state-of-the-art style transfer effect. All baselines were evaluated using publicly available implementations and default configurations. Furthermore, we compared our method with two advanced diffusion model-based style transfer methods, namely, InST and Z*. Our method adeptly transferred the tonality and brushstrokes of the style image while achieving the best content preservation effect. **Conclusion** Qualitative and quantitative comparisons with other arbitrary style transfer methods prove that our method can generate higher quality and more accurate style transfer results. Future work will focus on further optimizing the algorithm to enhance the model's adaptability to various styles and content, as well as exploring more innovative application scenarios. We anticipate that this method will play a more significant role in fields such as artistic creation and multimedia design, providing users with a richer and more personalized visual experience.

Key words: image generation; generative model; diffusion model; style transfer; attention mechanism

0 引言

一幅画作是艺术家创造力的结晶,其艺术创造

力体现在个性化的材质和笔触、捕捉的美丽瞬间以及特定的语义元素上。当试图利用一幅喜爱的画作创作新的数字艺术作品,以此模仿艺术家的原始创意时,所面临的任务便是示例引导的艺术图像生成。

基于一个或多个示例生成艺术图像的方法引起了广泛的研究兴趣。风格迁移(style transfer)技术的核心目标是开发一种风格转换机制,使得对于任何给定的艺术画作,人们能够根据其独特的风格对内容图像进行渲染,生成既保留内容图像的语义结构,又融入风格图像纹理特点的艺术化合成图像。任意风格迁移(arbitrary style transfer)技术进一步扩展了风格迁移的通用性和灵活性,其目标是通过一次性训练,获得一个能够处理各种不同风格的通用风格迁移模型。

在图像风格迁移的早期探索中,传统方法主要依赖于纹理合成和图像滤波技术,例如基于笔画的渲染和图像滤波。这些方法虽然在风格化效果上取得了一定成就,但它们通常受限于手工制作的低级特征,缺乏对图像整体语义的深入理解,导致在视觉效果和处理效率上存在局限。

随着深度学习技术的飞速发展,神经风格迁移(neural style transfer, NST)和生成对抗网络(generative adversarial network, GAN)为图像生成任务带来了新的视角。一些研究者将GAN应用于风格迁移领域,通过训练网络模型,实现了快速的风格化过程。随后,一系列编码器—解码器架构的方法相继提出,它们通过优化网络结构和损失函数设计,显著提高了风格迁移的实时性和效果。编码器—解码器架构已成为风格迁移领域的主流网络框架,它能够迅速提取图像的内容和风格特征。然而,由于其依赖于特征的二阶统计量进行风格度量,生成的图像有时会出现伪影等问题。另一方面,基于生成对抗网络的风格迁移方法虽然在提高生成图像逼真度方面取得了显著进展,但对抗损失的收敛问题以及模式崩溃和梯度消失等挑战仍然存在。

基于扩散模型(diffusion model)的生成方法因其多样化的生成结果和高质量的效果而受到广泛关注。扩散模型的核心原理是逐步向数据中添加噪声,直至数据变得不可辨识,随后通过逆过程逐步去除噪声,恢复数据的原始状态。与传统风格迁移框架相比,扩散模型在生成图像的多样性、质量和逼真度方面展现出明显优势,并能有效捕捉数据的复杂分布。扩散模型因其强大的数据建模能力,特别适合于多模态生成任务。例如,在将自然语言描述转化为图像的领域,扩散模型能够理解文本的语义内容,并生成与之相匹配的高质量图像。

将文字到图像生成的扩散模型应用于风格迁移任务是合情合理的,但这项技术仍面临一些挑战。首先,风格迁移所涉及的艺术风格通常具有高度的视觉和艺术特征,这些特征很难通过语言精确表达。尽管文本引导的风格化技术能够利用自然语言和文本提示生成艺术图像,但目标风格的文本提示往往仅限于对材料、艺术流派、艺术家或著名艺术作品的粗略描述。为了再现生动的内容和风格,通常需要详细的辅助文本输入来引导生成过程。即便如此,生成的结果可能仍难以完全捕捉到特定绘画作品的创意和构思。研究者在图像生成领域探索了使用参考图像作为视觉提示的方法,以提升生成图像的质量和多样性。不过,这些技术通常需要对扩散模型进行额外训练,伴随着高昂的微调成本,包括对大量数据和计算资源的需求。

为了克服这些方法对计算资源和时间的高要求,本文提出一种创新的无需额外训练的风格迁移方法,既利用了扩散模型高质量生成的优势,又保持了低计算成本。本文方法利用摄影图像和艺术图像作为视觉提示注入扩散模型网络中,以提供参考信息指导新图像生成。具体而言,本文方法利用扩散模型中的注意力机制,将参考图信息通过一种新颖的注意力交换算法注入图像生成路径中。本文方法无需微调模型,保持了足够的灵活性和高效性。

本文主要贡献归纳如下:1)提出一种基于扩散模型的风格迁移方法,利用扩散模型的优势,解决了以往方法中的伪影和风格多样性不足的问题;2)所提出的风格迁移方法无需微调和额外训练,通过聚焦于扩散模型推理过程中的注意力机制,设计了一种新颖的三线并行结构和一种引导信息注入机制,减少了对计算资源和时间的需求;3)通过与其他任意风格迁移方法的定性和定量比较,证明了本文方法能够生成更高质量、风格更准确风格迁移结果。

1 相关工作

1.1 图像风格迁移

传统的风格迁移方法,如基于笔触的渲染和图像过滤通常使用低级手工制作的特征。随着深度学习技术的飞速发展,特别是深度卷积神经网络(convolutional neural network, CNN)的应用, Gatys 等人

(2016)首次提出神经风格迁移,这是一种创新的方法,通过利用卷积网络提取的特征图之间的格拉姆(Gram)矩阵来表征风格。Li等人(2017)进一步提出一种使用逐通道间特征的均值和方差表示风格的方法。尽管这些基于CNN的方法在风格迁移领域取得了显著成果,但它们将任务表述为一个复杂的优化问题,从而导致了较高的计算成本,限制了实时生成风格化结果的能力。

Goodfellow等人(2020)提出的生成对抗网络为图像生成任务带来了新的视角。一些研究者将GAN应用于风格迁移领域,通过训练网络模型,实现了快速风格化过程(Zhu等,2017;Mirza和Osindero,2014)。一些方法依靠可学习的神经网络匹配特征空间中的统计信息以提高效率。单风格单模型的方法(Johnson等,2016)为每种风格训练一个特定的网络。每个模型多种风格的方法(Chen等,2021)使用一个模型表示多种风格。

任意风格迁移方法构建了更灵活的前馈架构,以使用统一模型处理任意风格。AdaIN(adaptive instance normalization)(Huang和Belongie,2017)直接将内容特征的总体统计数据与风格特征的统计数据对齐,并采用条件实例规范化。Li等人(2019)学习跨域特征线性变换矩阵以实现通用风格转换,并通过从转换后的特征解码生成所需的风格化结果。Park和Lee(2019)引入SANet(style-attentional networks)以灵活地将语义上最近的风格特征匹配到内容特征上。An等人(2021)提出可逆神经流和无偏特征传输模块(ArtFlow)以防止通用风格转换期间的內容泄漏。Liu等人(2021)提出一个自适应注意力规范化模块(adaptive attention normalization, AdaAttN),以同时考虑浅层和深层特征来计算注意力分数。Deng等人(2022)提出名为StyTr2(image style transfer with Transformers)的风格迁移方法,实现基于视觉Transformer的无偏风格迁移。Zhang等人(2022)提出基于对比学习的风格表示和风格迁移方式。Hong等人(2023)提出审美模式感知风格迁移网络(aesthetic pattern-aware style transfer networks, AesPA-Net)。基于GAN的方法已成功用于集合风格迁移,这些方法将集合中的风格图像视为一个域。刘哲良等人(2019)提出一种结合全卷积网络(fully convolutional networks, FCN)与CycleGAN的图像风格迁移方法,使得图像能够实现特定目标之

间的实例风格迁移。廖远鸿等人(2023)提出一种能够降低失真并生成不同风格强度图像的人脸风格迁移算法MStarGAN(multilayer StarGAN)。随着扩散模型的发展,一些基于扩散模型的风格迁移方法相继提出。Zhang等人(2023)提出一种基于注意力的反转风格转移方法,称为InST(inversion-based style transfer),这是一种需要额外训练的方法。Deng等人(2023)在扩散模型中引入了一种交叉注意重排策略,称为Z*,这是一种在交叉注意力机制中实现风格迁移的方法。孙梅婷等人(2023)提出一种基于Transformer的任意风格迁移方法,求取内容特征与风格特征间的关联性,并将内容特征替换为风格特征的加权和,实现风格转换。

1.2 扩散模型

扩散模型作为一种生成模型,近年来在图像生成领域取得了重要进展。其基本思想是将数据逐步添加噪声直到完全随机化,然后通过逐步去噪恢复原始数据。像素级别的扩散模型主要思想是将图像像素逐步添加高斯噪声,并训练一个去噪网络来逆向恢复图像。Ho等人(2020)提出的去噪扩散概率模型(denoising diffusion probabilistic model, DDPM)通过改进这一过程,使得扩散模型在生成质量和训练效率上均有所提升。Rombach等人(2022)提出的潜变量扩散模型(latent diffusion model, LDM)是扩散模型的一个重要改进方向。与像素级别的扩散模型不同,LDM在潜变量空间中进行扩散和逆向过程。LDM降低了计算复杂度,提高了生成效率。

研究者在图像生成领域探索了使用参考图像作为视觉提示的方法,以提升生成图像的质量和多样性。这些方法主要包括:一些研究(Ruiz等,2023;Kumari等,2023)通过包含相同风格的图像集微调扩散模型实现风格迁移。另有研究(Gal等,2022;Avrahami等,2022)基于文本嵌入技术,学习用户给定的图像风格,实现风格迁移技术。随着扩散模型的发展,一些基于扩散模型的风格迁移方法相继提出。Zhang等人(2023)提出一种反演风格转移方法。上述技术通常伴随着高昂的微调成本,包括对大量数据和计算资源的需求。

1.3 注意力机制

注意力机制是深度学习中的一种关键技术,特别是在序列模型和图像处理领域中。它允许模型在处理数据时,能够聚焦于输入数据中最重要的部分。

注意力机制的多种变体为模型提供了不同的方式捕捉和处理数据中的相关信息,其中与扩散模型相关的注意力机制包括自注意力(self-attention)、多头注意力(multi-head attention)和交叉注意力(cross-attention)机制(Vaswani等, 2017)。Park和Lee (2019)将注意力机制引入风格迁移,提出SANet以灵活地将语义最相似的风格特征匹配到内容特征上。这一方法基于传统编解码器结构,难以保证较高的生成质量。Deng等人(2023)通过交叉注意重排策略,提出一种风格迁移方法。这一方法虽然引入了注意力机制,但没有关注到自注意力机制在扩散模型中的重要作用。

2 方法

2.1 预先知识

2.1.1 潜在扩散模型

扩散概率模型是一种概率模型,旨在通过逐步去噪一个正态分布变量来学习数据分布,这对应于学习一个固定马尔可夫链长度为 T 的逆过程。对于图像合成,模型可以解释为等权重的去噪自编码器序列 $\epsilon_\theta(\mathbf{x}_t, t)$; $t = 1, \dots, T$,这些自编码器被训练来预测其输入 $\mathbf{x}_t$ 的去噪变体,其中 $\mathbf{x}_t$ 是输入 $\mathbf{x}$ 的噪声版本。对于随机扩散模型,给定均值估计器 μ_T 和 $\mathbf{x}_T \sim \mathcal{N}(0, I)$,图像 $\mathbf{x} = \mathbf{x}_0$ 是通过 $\mathbf{x}_{1:T} \sim \mathcal{N}(\mu_T(\mathbf{x}_t, t), \text{diag}(\sigma_t^2))$ 生成的。确定性扩散模型使用常微分方程式生成图像。给定均值估计器 μ_T ,确定性扩散模型通过 $\mathbf{z} = \mathbf{x}_T \sim \mathcal{N}(0, I)$ 生成 $\mathbf{x} = \mathbf{x}_0$ 。潜在扩散模型首先使用一个自编码模型,该模型学习一个与图像空间在感知上等价的潜在空间,以计算潜在代码。编码器 E 将图像 $\mathbf{x}$ 映射到空间潜在代码 $\mathbf{z} = E(\mathbf{x})$ 。解码器 D 学习将这些潜在代码映射回图像,使得 $D(E(\mathbf{x})) \approx \mathbf{x}$ 。在学到的潜在空间中训练一个扩散模型以生成目标潜在代码。条件LDM损失计算为

$$L_{\text{LDM}} = \mathbb{E}_{E(\mathbf{x}), \epsilon \sim \mathcal{N}(0, I), t} \left\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t) \right\|_2^2 \quad (1)$$

式中, $\mathbb{E}$ 为期望, θ 表示模型参数, t 是时间步骤, $\mathbf{z}_t$ 是该时刻的潜在表示。

潜在扩散模型的基础UNet骨干网络通过交叉注意力机制得到增强,这对于学习各种输入模态(例如,类别标签、分割掩码和联合训练的文本嵌入)的基于注意力的模型非常有效。将条件输入 y 转换为

条件向量的模型称为 $\tau_\theta(y)$ 。然后,条件LDM损失被确定,具体为

$$L_{\text{LDM}} = \mathbb{E}_{E(\mathbf{x}), y, \epsilon \sim \mathcal{N}(0, I), t} \left\| \epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y)) \right\|_2^2 \quad (2)$$

2.1.2 多头自注意力机制

注意力机制的核心思想是让模型在处理数据时能够动态地关注输入序列中的特定部分。这通常通过一个可学习的权重向量实现,该向量能够对输入序列的每个元素分配不同的关注程度。注意力机制通常通过一个权重向量来实现,这个权重向量是模型在训练过程中学习得到的。对于序列数据,每个元素都会根据其当前任务的相关性获得一个权重值。这些权重值通常通过一个可学习的函数计算得出,该函数会考虑输入元素与当前上下文的关系。

在扩散模型中,多头自注意力机制是一种关键技术,它允许模型在生成数据的过程中,同时考虑输入数据的多个不同方面。首先,将输入序列映射到多个不同的表示空间中,每个表示空间对应一个注意力头。对于每个注意力头,分别生成查询(Q)、键(K)和值(V)矩阵。这些矩阵是通过将输入表示与不同的可学习权重矩阵相乘得到的。对于每个头,计算查询矩阵与键矩阵的点积,然后除以缩放因子(通常是键向量维度的平方根),以防止梯度消失或爆炸。

$$\text{Score}_{i,j} = \frac{Q_i K_j^T}{\sqrt{d_k}} \quad (3)$$

式中, Q_i 和 K_j 分别是第 i 个查询和第 j 个键, d_k 是键向量的维度。使用softmax函数对每个头的注意力分数进行归一化,得到注意力权重为

$$W_i = f_{\text{softmax}}(S_i) \quad (4)$$

将每个头的值矩阵与其对应的归一化权重相乘,然后对所有头的结果进行求和,得到最终的输出表示,具体为

$$O_{\text{output}} = \sum_{i=1}^h W_i \times V_i \quad (5)$$

式中, h 是注意力头的数量。多头自注意力机制在扩散模型中的应用,使得模型能够更有效地捕捉和生成数据的复杂特征,从而提高生成数据的质量和多样性。

2.2 概述

本文提出的风格迁移方法的流程图如图1所示,用户给定一幅目标内容图和一幅目标艺术图,首先经过特征提取网络获得图像特征,然后经过

DDIM (denoising diffusion implicit model) 反演算法 (Song 等, 2022) 获得内容特征集和风格特征集。内容特征集、风格特征集和初始特征被共同送入生成网络进行一种新颖的三线并行图像生成过程, 旨在同时处理内容特征、风格特征和融合过程, 以实现更高效和高质量的风格迁移效果。该三线并行过程如图 2 所示, 涵盖了图像生成路径、内容引导路径和风格

引导路径。核心的生成路径涵盖了从初始噪声状态到最终图像的完整转化流程。内容引导路径专注于从内容图像中提取关键信息, 并将其融合入生成过程中的特定内容模块。相应地, 风格引导路径致力于从风格图像中抽取特征, 并将其嵌入到生成过程中的特定风格模块, 以此达成双重图像引导的风格迁移目标。

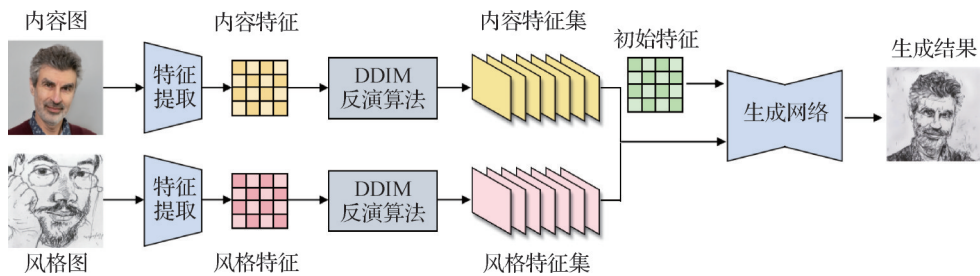


图 1 本文方法的流程图
Fig. 1 The pipeline of proposed method

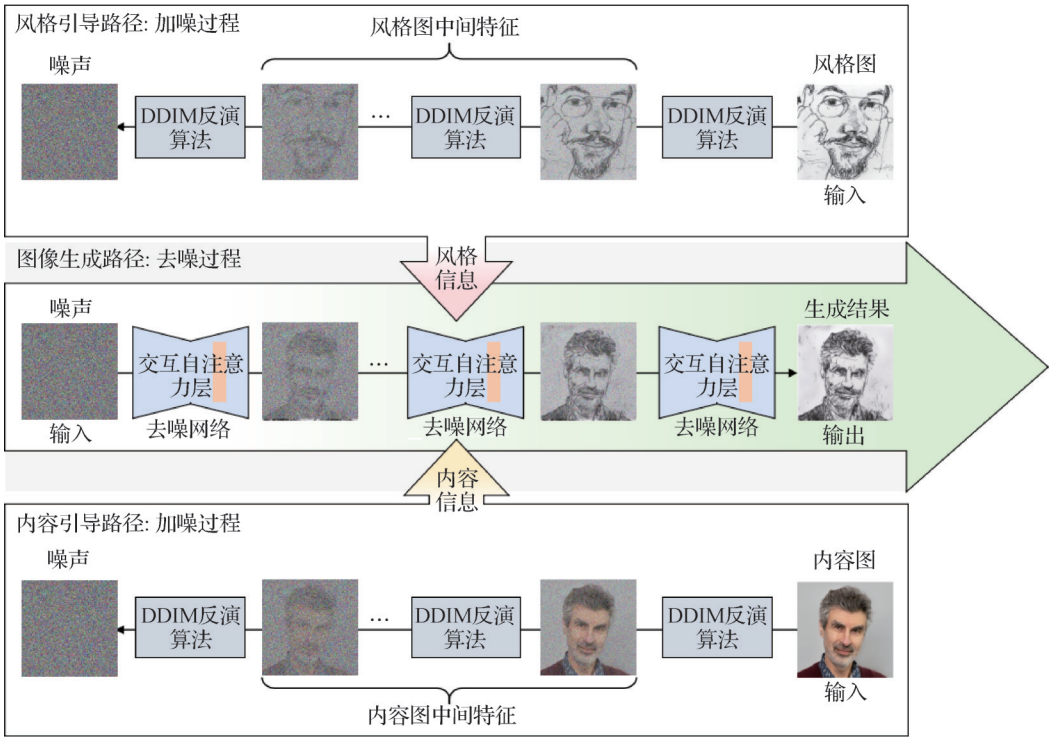


图 2 本文提出的三线并行风格迁移框架
Fig. 2 The three-branch parallel style transfer framework proposed in this paper

本文方法基于一个预训练的潜在扩散模型, 该模型由编码器、去噪网络和解码器 3 个主要部分组成。如图 3 所示, 在这一预训练模型的基础上, 进一步设计了交互注意力模块和特征对齐模块。交互注意力模块通过将参考图像的特征定义为自注意力机制中的键(K)和值(V), 同时将生成图像的结构信息

定义为查询(Q), 实现了风格和-content 图像信息的有效融合, 不仅保留了内容的完整性, 也实现了风格上的迁移。此外, 特征对齐模块负责计算风格图像与生成图像在潜在特征空间的统计差异, 通过调整生成图像特征的均值和方差与风格图像相匹配, 进一步提高了风格迁移的精确度。

2.3 风格迁移网络

2.3.1 交互注意力模块

图像风格迁移的核心有两点,即迁移艺术图像风格,同时保持内容图像内容信息。本文设计了新颖的三线并行生成路径,即新图像生成路径、内容参考路径和风格参考路径,目标是将风格和-content特征融入图像生成过程中。这种特征融合通过交互注意力模块实现。具体而言,这一过程始于3个生成路径:一条从文本提示和初始噪声出发,自然形成与文本提示相匹配的图像 I ;另外两条则分别导向一个艺术图像 I_s 和一个内容图像 I_c 。本文使用DDIM反演(Song等,2022)对参考图像进行加噪,获得参考图去噪路径上的中间状态。

为了生成融合了参考风格特征的图像 I ,利用了扩散模型的层次化生成特性,即在模型深层网络生成结构化信息,浅层网络生成纹理、颜色等信息,

这两者分别对应任务设定中的内容和风格信息。如图3所示,基于这种特性,在扩散模型主干网络的深层自注意力模块中,将原始路径中的关键和值特征与内容路径中相应的 K_c 和 V_c 进行交换。随后,在扩散模型主干网络的浅层自注意力模块中,将原始路径中的关键和值特征与风格路径中相应的 K_s 和 V_s 进行交换。这样,注意力机制利用原始查询特征 Q_l 与参考关键特征 K_c 、 K_s 之间的相似度,以权重形式重构参考值特征。以上交互注意力过程表示为

$$f_{\text{Attention}}(Q_l, K_c, V_c) = f_{\text{softmax}}\left(\frac{Q_l K_c^T}{\sqrt{d}}\right) V_c \quad (6)$$

$$f_{\text{Attention}}(Q_l, K_s, V_s) = f_{\text{softmax}}\left(\frac{Q_l K_s^T}{\sqrt{d}}\right) V_s \quad (7)$$

通过这种交换自注意力编码的方式,内容和图像信息以 K 、 V 矩阵的形式注入新图像生成过程。

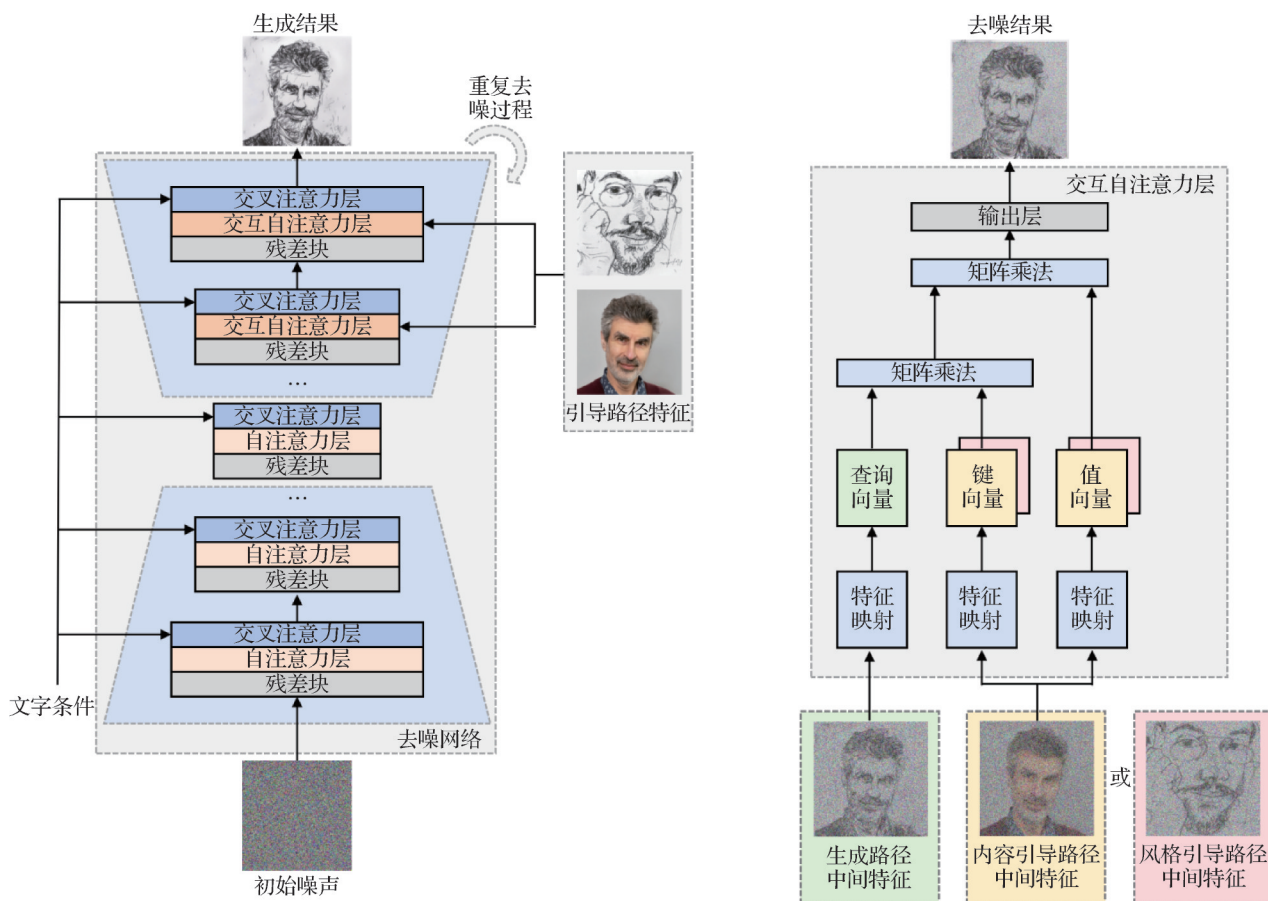


图3 本文交互注意力模块示意图

Fig. 3 The schematic diagram of the swapping attention module in this paper

2.3.2 图像特征对齐模块

色调一致性是评价风格迁移效果的关键因素。

尽管基于交互注意力的方法在转移艺术图像中的笔触和纹理方面表现出色,但它们通常难以实现内容

图像的整体色调与风格图像的对齐。扩散模型的图像生成不仅依赖于生成网络和文本条件,还受到初始噪声的显著影响。先前的研究已经指出,生成图像的色彩基调与初始噪声的采样紧密相关。基于这一发现,本文通过调整初始噪声来解决生成图像与风格图像之间的色调一致性问题。

本文提出一种新的方法,通过将生成图像的初始噪声与风格图像的分布对齐,提高色调一致性。这种方法与传统风格迁移中的 AdaIN 计算方式相似。具体来说,对艺术图像和内容图像分别进行加噪处理。接着,为了整合两个初始潜在噪声中的关键信息,利用 AdaIN 技术进行调整,以有效传递色调信息。具体的数学表达式为

$$z_T = \sigma(z_T^c) \left(\frac{z_T^c - \mu(z_T^c)}{\sigma(z_T^c)} \right) + \mu(z_T^s) \quad (8)$$

式中, $\mu(\cdot)$ 表示通道均值, $\sigma(\cdot)$ 表示通道标准差, T 表示时间步, z_T^s 表示风格图像潜在特征, z_T^c 表示内容图像潜在特征。通过这种方式,确保了生成图像潜在特征 z_T 在保持内容图像结构的同时,其色调与风格图像的色调相匹配。

3 实验结果与分析

3.1 实验设计

实验与几种传统风格迁移方法进行比较,包括 ArtFlow (An 等, 2021)、AdaAttN (Liu 等, 2021)、StyTr2 (Deng 等, 2022)、CAST (contrastive arbitrary style transfer) (Zhang 等, 2022) 和 AesPA-Net (aesthetic pattern-aware style transfer networks) (Hong 等, 2023)。所有的基线 (Hong 等, 2023) 都是使用公开可用的实现和默认配置进行测试的。

测试数据集方面,使用了收集自维基艺术 (WikiArt) (Phillips 和 Mackintosh, 2011) 的 100 幅艺术图像和来自 Places365 (Zhou 等, 2018) 数据集的 100 幅真实图像,这些图像分别从 10 万级别的数据集中随机采样得到,以达到公平比较的结果。

实现细节上,在所有实验中,使用 SDXL (stable diffusion extra large) (Podell 等, 2023), 并使用默认的超参数;使用具有扩散步骤和引导比例的 DDIM 采样器;使用 SDXL 中的默认的文本编码器。在 NVIDIA GeForce RTX4090 上,每幅图像的生成过程

大约需要 30 s, 批量大小为 1。

3.2 对比实验

本研究采用两种主要的性能评估指标衡量本文的风格迁移方法: LPIPS (learned perceptual image patch similarity) 和欺骗率。LPIPS 是一种衡量图像之间感知相似度的指标,它基于深度学习模型 VGG-19 (Visual Geometry Group) 的特征提取能力。LPIPS 是一种先进的图像质量评估指标,它通过深度学习技术模拟人眼对图像的感知能力。这一指标利用预训练的 CNN 提取原始图像和目标图像的特征表示,然后通过计算这些特征的感知距离来评估两幅图像的相似性。LPIPS 得分越低,表明重建图像在视觉上与原始图像的相似度越高,质量越好。LPIPS 特别适用于图像风格迁移、图像生成和图像编辑等领域,因其能够更准确地反映人眼对图像质量的感知。LPIPS 计算式为

$$d(\mathbf{x}, \mathbf{x}_0) = \sum_{l=1}^L \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \mathbf{w}_l \odot (\hat{\mathbf{y}}_{hw}^l - \hat{\mathbf{y}}_{0hw}^l) \right\|_2^2 \quad (9)$$

式中, $d(\mathbf{x}, \mathbf{x}_0)$ 表示两幅图像 $\mathbf{x}$ 和 $\mathbf{x}_0$ 之间的距离, L 表示图像特征的层级数或尺度数, l 表示特征图层数, H 和 W 分别表示特征图的高度和宽度, w 表示权重因子,用于平衡不同层级的贡献, $\hat{\mathbf{y}}_{hw}^l$ 表示在第 l 层、位置 (h, w) 的特征响应, $\hat{\mathbf{y}}_{0hw}^l$ 表示在第 l 层、位置 (h, w) 的参考图像的特征响应, $\odot$ 表示逐元素乘法, $\|\cdot\|_2$ 表示 L2 范数。

欺骗率用于衡量风格迁移结果的逼真度,即风格化图像能否被一个训练有素的分类器正确识别为目标风格。这个指标反映了风格迁移方法在模仿特定艺术风格方面的有效性。欺骗率高意味着风格迁移结果在视觉上更接近目标风格,足以误导分类器。令 C 是分类器, S 是目标风格集合, $\hat{\mathbf{y}}$ 是分类器对风格化图像 $\mathbf{y}$ 的预测结果, y_{true} 是 $\mathbf{y}$ 的真实风格标签。欺骗率可以表示为

$$f_{\text{DR}} = \frac{1}{M} \sum_{j=1}^M [C(y_j) = y_{\text{true},j}] \quad (10)$$

式中, M 是风格化图像的数量, $[C(y_j) = y_{\text{true},j}]$ 是一个指示函数,当分类器的预测结果等于真实风格标签时,该函数值为 1, 否则为 0。

定量实验结果如表 1 所示,本文方法取得了最高的欺骗率,同时 LPIPS 得分也较好,证明了本文方法在风格准确度和内容保持上的先进性。

3.3 结果展示与分析

3.3.1 与传统风格迁移方法比较

与 5 种传统风格迁移方法进行比较, 包括 ArtFlow、AdaAttN、StyTr2、CAST 和 AesPA-Net。如图 4 所示, 大部分基线方法能够较好地迁移颜色和笔触, 但是难以迁移具有一定形状的风格, 如第 2 行风格图中的树木, 以及较为有特色的风格中特殊技法, 如第 4、7、8 行的水墨风格中的软笔笔触和留白的特点。此外, AesPA-Net 在第 4 行的天空中生成了较为明显的伪影, 第 8 行的结果中内容丢失较为明显。CAST 在画作真实感和水墨风格中取得了较好的表现, 但是在第 6 行的天空中生成了一些伪影。StyTr2 的结果存在模糊、伪影的问题, 如第 3—5 行。AdaAttN 能够较好地保持内容, 但是这影响了该方

表 1 6 种风格迁移方法生成质量对比

Table 1 The comparison of generation quality among six style transfer methods

方法	LPIPS↓	欺骗率/%↑
AesPA-Net(Hong 等, 2023)	0.318	62.80
CAST(Zhang 等 2022)	0.315	<u>64.20</u>
StyTr2(Deng 等, 2022)	0.310	54.20
AdaAttN(Liu 等, 2021)	0.314	51.60
ArtFlow(An 等, 2021)	<u>0.311</u>	44.20
本文	0.313	68.40

注: 加粗字体表示各列最优结果。↑表示数值越高越好; ↓表示数值越低越好。

法的风格表现力, 如第 3、6 行中难以复现风格图中的笔触感。ArtFlow 内容保持得更加完整, 但同样存

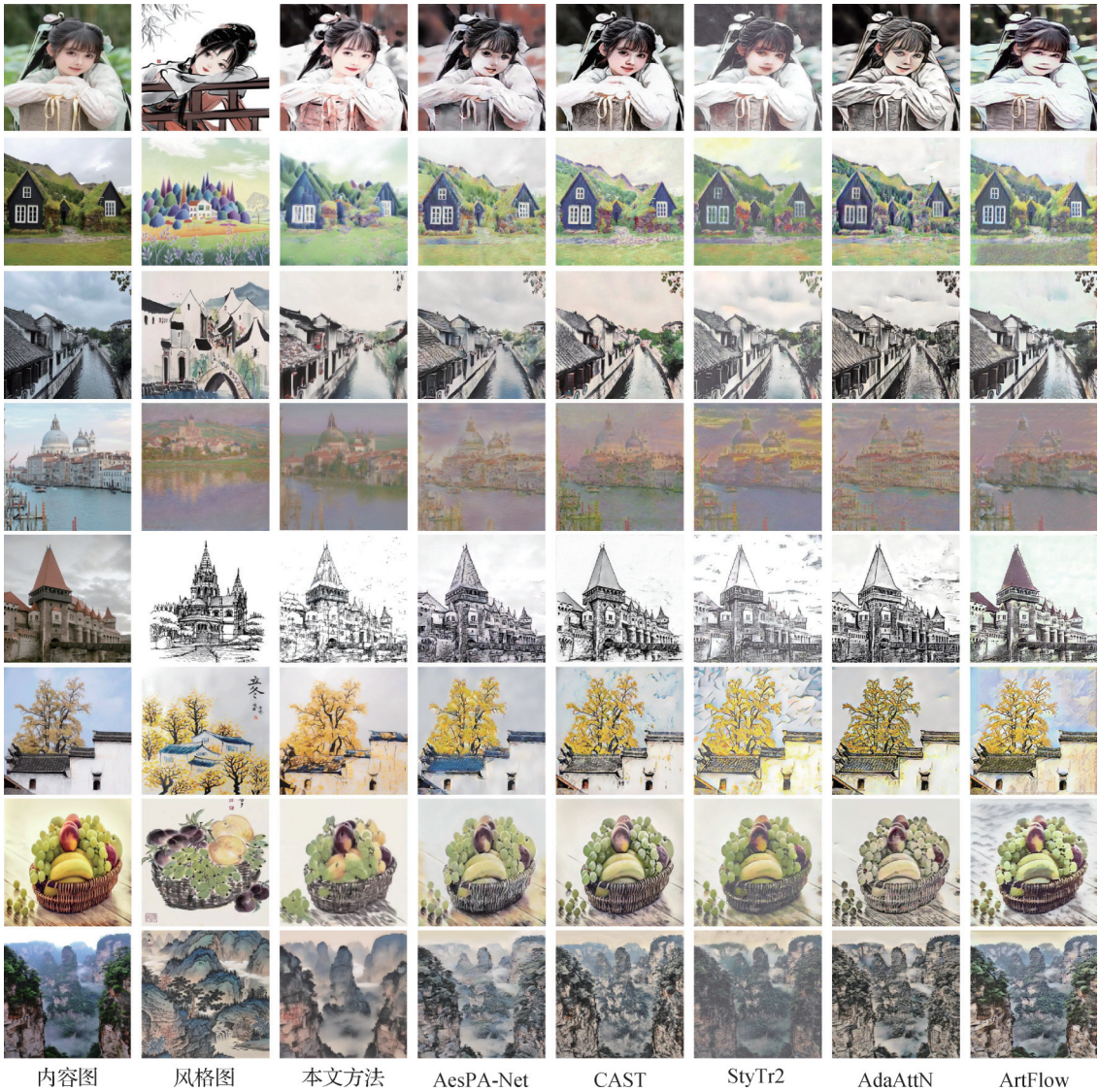


图 4 与传统风格迁移方法的对比图

Fig. 4 The comparison of traditional style transfer methods

在风格不准确和伪影的问题。相较之下,本文方法能够捕捉风格图像中多层次的风格信息,忠实地迁移色调、笔触和形状等多方面的艺术风格,取得了最优秀的效果。

3.3.2 与基于扩散模型的方法比较

与两种先进的基于扩散模型的风格迁移方法 InST 和 Z* 进行比较。如图 5 所示,本文方法能够很好地迁移风格图的色调和笔触,同时取得最好的内容保持效果。具体而言,InST 是一种需要额外训练的风格迁移方法,能够从语义层次迁移图像属性,因此能够最大程度迁移风格图的信息,但它难以迁移风格图的色调信息,并一定程度丢失了内容图信息。Z* 同样难以迁移风格图的色调,这削弱了该方法的风格表现力。相较之下,本文方法取得了最先进的风格迁移效果。

与基于扩散模型的图像处理方法类似,本文方法在语义对应的风格图和内容图对上能取得更好的效果。

3.4 消融实验

图 6 分别展示了消融方法中的交互注意力模块、特征对齐(AdaIN)模块,以及完整模型风格迁移的结果,以分析不同模块的作用。如图 6(c)所示,去除交互注意力模块后,模型虽然在 AdaIN 模块的作用下能够迁移风格图的整体色调,但是难以复现不同的笔触等风格信息。如图 6(d)所示,加入交互



图 5 与基于扩散模型的风格迁移的对比图

Fig. 5 The comparison of style transfer based on diffusion models

注意力模块能够很好地迁移笔触,但由于没有 AdaIN 模块辅助模型迁移整体色调,生成图像的色调与内容图更相近。如图 6(e)所示,完整模型能够生成更准确、更高质量的结果。

3.5 局限性

尽管所提出的风格迁移方法在实现艺术化图像转换方面取得了显著成效,但仍面临若干局限性。首先,推理过程的速率受到了一定影响,原因在于本文方法在推理时需进行注意力图的交换操作,该操作引入了额外的计算量。鉴于本文方法避免了耗时的模型微调过程,其时间与计算成本相较于那些需要微调参数的模型依然低得多,且图像生成速度保持在 30 s 以内,这在实际应用中是可被广泛接受的。



图 6 消融实验结果

Fig. 6 Ablation study results

((a)content images;(b)style images;(c)without interactive attention module;(d)without AdaIN module;(e)full model)

其次,对于某些风格,尤其是那些要求对主体形状进行显著变化的风格,例如夸张漫画风格,本文方法可能需要对风格注入层进行特别调整,以便能够实现预期的剧烈形变效果。对于存在剧烈的形状变化的风格,如抽象主义绘画,本文方法可能存在失败案例。这表明,尽管本文方法在多种风格迁移任务中表现出色,但在处理某些特殊风格时,仍需进一步的优化和调整,以达到最佳效果。

4 结 论

针对传统风格迁移中图像质量和多样性的挑战,以及先进的扩散模型中存在的风格表达不准确、计算和时间成本高的问题,本文提出一种新颖的三线并行风格迁移框架。该框架通过图像生成路径、内容引导路径和风格引导路径的协同工作,实现了从初始噪声到目标图像的高效转化。本文方法基于预训练的潜在扩散模型,引入了交互注意力模块和特征对齐模块。交互注意力模块利用参考图像的特征作为自注意力机制的键和值,同时以生成图像的结构信息作为查询,实现了内容与风格的有效融合。特征对齐模块则通过计算风格图像与生成图像在潜在特征空间的统计差异,调整生成图像的特征以匹配风格图像,从而提高了风格迁移的精确度。未来工作将着眼于进一步优化算法,提升模型对不同风格 and 内容的适应性,以及探索更多创新的应用场景。期望这一方法能够在艺术创作、多媒体设计等领域发挥更大的作用,为用户带来更加丰富和个性化的视觉体验。

参考文献

- An J, Huang S Y, Song Y B, Dou D J, Liu W and Luo J B. 2021. ArtFlow: unbiased image style transfer via reversible neural flows//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 862-871 [DOI: 10.1109/CVPR46437.2021.00092]
- Avrahami O, Lischinski D and Fried O. 2022. Blended diffusion for text-driven editing of natural images//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18187-18197 [DOI: 10.1109/CVPR52688.2022.01767]
- Chen H B, Zhao L, Wang Z Z, Zhang H M, Zuo Z W, Li A L, Xing W and Lu D M. 2021. DualAST: dual style-learning networks for artistic style transfer//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 872-881 [DOI: 10.1109/CVPR46437.2021.00093]
- Deng Y Y, He X Y, Tang F and Dong W M. 2023. Z*: zero-shot style transfer via attention rearrangement [EB/OL]. [2024-10-15]. <https://arxiv.org/pdf/2311.16491.pdf>
- Deng Y Y, Tang F, Dong W M, Ma C Y, Pan X J, Wang L and Xu C S. 2022. StyTr²: image style transfer with transformers//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11316-11326 [DOI: 10.1109/CVPR52688.2022.01104]
- Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano A H, Chechik G and Cohen-Or D. 2022. An image is worth one word: personalizing text-to-image generation using textual inversion [EB/OL]. [2024-10-15]. <https://arxiv.org/pdf/2208.01618.pdf>
- Gatys L A, Ecker A S and Bethge M. 2016. Image style transfer using convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2414-2423 [DOI: 10.1109/cvpr.2016.265]
- Goodfellow I, Pouget A J, Mirza M, Xu B, Warde F D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM, 63 (11): 139-144. [DOI: 10.1145/3422622]
- Ho J, Jain A and Abbeel P. 2020. Denoising diffusion probabilistic models//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 6840-6851
- Hong K, Jeon S, Lee J, Ahn N, Kim K, Lee P, Kim D, Uh Y and Byun H. 2023. AesPA-Net: aesthetic pattern-aware style transfer networks//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 22701-22710 [DOI: 10.1109/ICCV51070.2023.02080]
- Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/ICCV.2017.167]
- Johnson J, Alahi A and Li F F. 2016. Perceptual losses for real-time style transfer and super-resolution//Proceedings of the 14th European Conference on Computer Vision—ECCV 2016. Amsterdam, the Netherlands: Springer: 694-711 [DOI: 10.1007/978-3-319-46475-6_43]
- Kumari N, Zhang B L, Zhang R, Shechtman E and Zhu J Y. 2023. Multi-concept customization of text-to-image diffusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1931-1941 [DOI: 10.1109/cvpr52729.2023.00192]
- Li X T, Liu S F, Kautz J and Yang M X. 2019. Learning linear transformations for fast image and video style transfer//Proceedings of 2019

- IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3809-3817 [DOI: 10.1109/CVPR.2019.003934]
- Li Y J, Fang C, Yang J M, Wang Z W, Lu X and Yang M H. 2017. Universal style transfer via feature transforms//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 385-395
- Liao Y H, Qian W H and Cao J D. 2023. MStarGAN: a face style transfer network with changeable style intensity. *Journal of Image and Graphics*, 28(12): 3784-3796 (廖远鸿, 钱文华, 曹进德. 2023. 风格强度可变的人脸风格迁移网络. *中国图象图形学报*, 28(12): 3784-3796) [DOI: 10.11834/jig.221149]
- Liu S H, Lin T W, He D L, Li F, Wang M L, Li X, Sun Z X, Li Q and Ding E R. 2021. AdaAttN: revisit attention mechanism in arbitrary neural style transfer//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 6629-6638 [DOI: 10.1109/ICCV48922.2021.00658]
- Liu Z L, Zhu W and Yuan Z Y. 2019. Image instance style transfer combined with fully convolutional network and cycleGAN. *Journal of Image and Graphics*, 24(8): 1283-1291 (刘哲良, 朱玮, 袁梓洋. 2019. 结合全卷积网络与 CycleGAN 的图像实例风格迁移. *中国图象图形学报*, 24(8): 1283-1291) [DOI: 10.11834/jig.180624]
- Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2024-10-15]. <https://arxiv.org/pdf/1411.1784.pdf>
- Park D Y and Lee K H. 2019. Arbitrary style transfer with style-attentional networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5873-5881 [DOI: 10.1109/CVPR.2019.00603]
- Phillips F and Mackintosh B. 2011. Wiki art gallery, Inc.: a case for critical thinking. *Issues in Accounting Education*, 26(3): 593-608 [DOI: 10.2308/iace-50038]
- Podell D, English Z, Lacey K, Blattmann A, Dockhorn T, Müller J, Penna J and Rombach R. 2023. SDXL: improving latent diffusion models for high-resolution image synthesis [EB/OL]. [2024-10-15]. <https://arxiv.org/pdf/2307.01952.pdf>
- Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]
- Ruiz N, Li Y Z, Jampani V, Pritch Y, Rubinstein M and Aberman K. 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 22500-22510 [DOI: 10.1109/cvpr52729.2023.02155]
- Song J M, Meng C L and Ermon S. 2022. Denoising diffusion implicit models [EB/OL]. [2024-10-15]. <https://arxiv.org/pdf/2010.02502.pdf>
- Sun M T, Dai L Q and Tang J H. 2023. Transformer-based multi-style information transfer in image processing. *Journal of Image and Graphics*, 28(11): 3536-3549 (孙梅婷, 代龙泉, 唐金辉. 2023. 基于 Transformer 方法的任意风格迁移策略. *中国图象图形学报*, 28(11): 3536-3549) [DOI: 10.11834/jig.211237]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Zhang Y X, Huang N S, Tang F, Huang H B, Ma C Y, Dong W M and Xu C S. 2023. Inversion-based style transfer with diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 10146-10156 [DOI: 10.1109/CVPR52729.2023.00978]
- Zhang Y X, Tang F, Dong W M, Huang H B, Ma C Y, Lee T Y and Xu C S. 2022. Domain enhanced arbitrary image style transfer via contrastive learning//Proceedings of 2022 ACM SIGGRAPH Conference Proceedings. Vancouver, Canada: ACM: #12 [DOI: 10.1145/3528233.3530736]
- Zhou B L, Lapedriza A, Khosla A, Oliva A and Torralba A. 2018. Places: a 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452-1464 [DOI: 10.1109/tpami.2017.2723009]
- Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: doi.org/10.1109/iccv.2017.244]

作者简介

张宇欣,女,博士研究生,主要研究方向为图像与视频生成。E-mail:zhangyuxin2020@ia.ac.cn

董未名,通信作者,男,研究员,主要研究方向为可视媒体合成与计算艺术。E-mail:weiming.dong@ia.ac.cn

徐常胜,男,研究员,主要研究方向为计算机视觉与多媒体计算。E-mail:changsheng.xu@ia.ac.cn