云南大学软件学院期末课程报告

Final Course Report
School of Software, Yunnan University

个人成绩

序号	学号	姓名	成绩
1	20211910089	曹伊凡	
2	20211050131	黄斌	
3	20211180016	陈时律	
4	20211120034	周轩哲	
5	20211060248	兰子豪	

学　　期:2024 春季学期

课程名称:人工智能创新实践

任课教师:郭竞

实践题目:图像风格迁移的发展探究

联系电话:

电子邮件:

完成提交时间：2024 年 6 月 2 日

团队成员分工表

成员	工作内容	工作量占比
曹伊凡	CAP-VSTNet 模型搭建；可视化界面制作；	20
黄斌	LoRA 模型训练；DEADiff 模型；	20
周轩哲	前期调研；文档撰写；训练测试数据的采集与拍摄；	20
兰子豪	文档撰写；前期调研；PPT 制作；	20
陈时律	CAP-VSTNet 模型训练；模型测试与评估；	20

云南大学软件学院 2024-2025 学年春季学期

《人工智能创新实践》成绩考核表

课程名称：人工智能创新实践 指导教师：郭竞 项目名称：图像风格迁移的发展探究
年级：2021 级 专业：人工智能 学号：20211910089 姓名：曹伊凡

指标内容	分值	指标内涵及评估标准				得分
		A	B	C	D	
构思与选题 （C，15 分）						
选题意义	5	意义重大	意义较大	意义一般,属于简单的开发	无意义	
技术路线的可行程度	10	合理可行 具体且有创新	合理可行,具体	基本合理可行	不够合理	
设计 （D，20 分）						
设计内容	10	内容非常丰富	内容较丰富	内容一般	内容欠缺	
解决的关键技术问题	10	准确，范围合适 重点突出	基本准确	部分关键	未抓住关键	
实现 （I，30 分）						
项目完成的技术水平（个人）	10	难度很大 达到较高水平	难度较大 超出一般水平	难度一般 达到普通水平	难度小 很容易实现	
小组成员的工作量	10	高出平均要求工作量 15%以上	高出平均要求工作量	达到平均要求工作量	低于平均要求工作量	
团队精神	10	团队合作精神强	合作情况良好	合作情况一般	合作不好	
运作 （O，15 分）						
达到预期目标的程度	15	组织优秀 完全达到	组织良好 较好达到	组织一般 基本达到	组织混乱 未能达到	
文档总结情况（20 分）						
文字表达	5	文字表达非常好	文字表达较好	文字表达一般	文字表达差 意思不明了	
文档制作	5	制作非常专业化	制作良好	制作一般	制作效果差	
报告质量	5	报告非常完整	报告比较完整	完整程度一般	报告不完整	
口头表达能力	5	整体效果很好	整体效果良好	整体效果一般	整体效果差	
总分：						
评语						

综合得分：_____（满分 100 分）

指导教师签名： 2024 年 5 月 29 日

云南大学软件学院 2024-2025 学年春季学期

《人工智能创新实践》成绩考核表

课程名称：人工智能创新实践 指导教师：郭竞 项目名称：图像风格迁移的发展探究
年级：2021 级 专业：人工智能 学号：20211050131 姓名：黄斌

指标内容	分值	指标内涵及评估标准				得分
		A	B	C	D	
构思与选题 （C，15 分）						
选题意义	5	意义重大	意义较大	意义一般,属于简单的开发	无意义	
技术路线的可行程度	10	合理可行 具体且有创新	合理可行,具体	基本合理可行	不够合理	
设计 （D，20 分）						
设计内容	10	内容非常丰富	内容较丰富	内容一般	内容欠缺	
解决的关键技术问题	10	准确，范围合适 重点突出	基本准确	部分关键	未抓住关键	
实现 （I，30 分）						
项目完成的技术水平（个人）	10	难度很大 达到较高水平	难度较大 超出一般水平	难度一般 达到普通水平	难度小 很容易实现	
小组成员的工作量	10	高出平均要求工作量 15%以上	高出平均要求工作量	达到平均要求工作量	低于平均要求工作量	
团队精神	10	团队合作精神强	合作情况良好	合作情况一般	合作不好	
运作 （O，15 分）						
达到预期目标的程度	15	组织优秀 完全达到	组织良好 较好达到	组织一般 基本达到	组织混乱 未能达到	
文档总结情况（20 分）						
文字表达	5	文字表达非常好	文字表达较好	文字表达一般	文字表达差 意思不明了	
文档制作	5	制作非常专业化	制作良好	制作一般	制作效果差	
报告质量	5	报告非常完整	报告比较完整	完整程度一般	报告不完整	
口头表达能力	5	整体效果很好	整体效果良好	整体效果一般	整体效果差	
总分：						
评语						

综合得分：_____（满分 100 分）

指导教师签名： 2024 年 5 月 29 日

云南大学软件学院 2024-2025 学年春季学期

《人工智能创新实践》成绩考核表

课程名称：人工智能创新实践 指导教师：郭竞 项目名称：图像风格迁移的发展探究
年级：2021 级 专业：人工智能 学号：20211180016 姓名：陈时律

指标内容	分值	指标内涵及评估标准				得分
		A	B	C	D	
构思与选题 （C，15 分）						
选题意义	5	意义重大	意义较大	意义一般,属于简单的开发	无意义	
技术路线的可行程度	10	合理可行 具体且有创新	合理可行,具体	基本合理可行	不够合理	
设计 （D，20 分）						
设计内容	10	内容非常丰富	内容较丰富	内容一般	内容欠缺	
解决的关键技术问题	10	准确，范围合适 重点突出	基本准确	部分关键	未抓住关键	
实现 （I，30 分）						
项目完成的技术水平（个人）	10	难度很大 达到较高水平	难度较大 超出一般水平	难度一般 达到普通水平	难度小 很容易实现	
小组成员的工作量	10	高出平均要求工作量 15%以上	高出平均要求工作量	达到平均要求工作量	低于平均要求工作量	
团队精神	10	团队合作精神强	合作情况良好	合作情况一般	合作不好	
运作 （O，15 分）						
达到预期目标的程度	15	组织优秀 完全达到	组织良好 较好达到	组织一般 基本达到	组织混乱 未能达到	
文档总结情况（20 分）						
文字表达	5	文字表达非常好	文字表达较好	文字表达一般	文字表达差 意思不明了	
文档制作	5	制作非常专业化	制作良好	制作一般	制作效果差	
报告质量	5	报告非常完整	报告比较完整	完整程度一般	报告不完整	
口头表达能力	5	整体效果很好	整体效果良好	整体效果一般	整体效果差	
总分：						
评语						

综合得分：_____（满分 100 分）

指导教师签名： 2024 年 5 月 29 日

云南大学软件学院 2024-2025 学年春季学期

《人工智能创新实践》成绩考核表

课程名称：人工智能创新实践 指导教师：郭竞 项目名称：图像风格迁移的发展探究
年级：2021 级 专业：人工智能 学号：20211120034 姓名：周轩哲

指标内容	分值	指标内涵及评估标准				得分
		A	B	C	D	
构思与选题 （C，15 分）						
选题意义	5	意义重大	意义较大	意义一般,属于简单的开发	无意义	
技术路线的可行程度	10	合理可行 具体且有创新	合理可行,具体	基本合理可行	不够合理	
设计 （D，20 分）						
设计内容	10	内容非常丰富	内容较丰富	内容一般	内容欠缺	
解决的关键技术问题	10	准确，范围合适 重点突出	基本准确	部分关键	未抓住关键	
实现 （I，30 分）						
项目完成的技术水平（个人）	10	难度很大 达到较高水平	难度较大 超出一般水平	难度一般 达到普通水平	难度小 很容易实现	
小组成员的工作量	10	高出平均要求工作量 15%以上	高出平均要求工作量	达到平均要求工作量	低于平均要求工作量	
团队精神	10	团队合作精神强	合作情况良好	合作情况一般	合作不好	
运作 （O，15 分）						
达到预期目标的程度	15	组织优秀 完全达到	组织良好 较好达到	组织一般 基本达到	组织混乱 未能达到	
文档总结情况（20 分）						
文字表达	5	文字表达非常好	文字表达较好	文字表达一般	文字表达差 意思不明了	
文档制作	5	制作非常专业化	制作良好	制作一般	制作效果差	
报告质量	5	报告非常完整	报告比较完整	完整程度一般	报告不完整	
口头表达能力	5	整体效果很好	整体效果良好	整体效果一般	整体效果差	
总分：						
评语						

综合得分：_____（满分 100 分）

指导教师签名： 2024 年 5 月 29 日

云南大学软件学院 2024-2025 学年春季学期

《人工智能创新实践》成绩考核表

课程名称：人工智能创新实践 指导教师：郭竞 项目名称：图像风格迁移的发展探究
年级：2021 级 专业：人工智能 学号：20211060248 姓名：兰子豪

指标内容	分值	指标内涵及评估标准				得分
		A	B	C	D	
构思与选题 （C，15 分）						
选题意义	5	意义重大	意义较大	意义一般,属于简单的开发	无意义	
技术路线的可行程度	10	合理可行 具体且有创新	合理可行,具体	基本合理可行	不够合理	
设计 （D，20 分）						
设计内容	10	内容非常丰富	内容较丰富	内容一般	内容欠缺	
解决的关键技术问题	10	准确，范围合适 重点突出	基本准确	部分关键	未抓住关键	
实现 （I，30 分）						
项目完成的技术水平（个人）	10	难度很大 达到较高水平	难度较大 超出一般水平	难度一般 达到普通水平	难度小 很容易实现	
小组成员的工作量	10	高出平均要求工作量 15%以上	高出平均要求工作量	达到平均要求工作量	低于平均要求工作量	
团队精神	10	团队合作精神强	合作情况良好	合作情况一般	合作不好	
运作 （O，15 分）						
达到预期目标的程度	15	组织优秀 完全达到	组织良好 较好达到	组织一般 基本达到	组织混乱 未能达到	
文档总结情况（20 分）						
文字表达	5	文字表达非常好	文字表达较好	文字表达一般	文字表达差 意思不明了	
文档制作	5	制作非常专业化	制作良好	制作一般	制作效果差	
报告质量	5	报告非常完整	报告比较完整	完整程度一般	报告不完整	
口头表达能力	5	整体效果很好	整体效果良好	整体效果一般	整体效果差	
总分：						
评语						

综合得分：_____（满分 100 分）

指导教师签名： 2024 年 5 月 29 日

目录

图像风格迁移的发展探究

摘要： 本文通过完整全面的梳理阐述图像风格迁移的发展历程，并从中分析出图像风格迁移技术所面临的三个核心挑战——风格与内容的平衡，用户可控性，训练数据限制。本文针对上述三个挑战，以此设计实验，分别学习研究 CAP-VSTNet 模型，DEADiff 模型，以及 LoRA 模型。深入浅出的分析三个模型在上述问题的作用机理，整体而完善的完成了本次实验，并取得了较为满意的效果。

关键词：图像风格迁移；残差网络；扩散模型；LoRA 模型

一、图像风格迁移

图像风格迁移是计算机视觉领域的研究热点，它是一种利用特定风格图像对指定图像进行重新映射的过程，目的在于将一种图像的风格、特征迁移到另一张图像上来，从而形成一张具有指定风格的新图像。

生成特定风格的图像是图像迁移的目的，但是实现目的的手段是多种多样的，尤其是随着深度学习技术的发展，其强大的特征提取能力为图像风格迁移带来了新的可能和更高效的实现方式。根据时间线，图像风格迁移大体可以分为传统的图像风格迁移和基于神经网络的图像风格迁移。

二、图像风格迁移的发展历程

2.1 基于传统的风格迁移

传统的图像风格迁移可以从两个角度来考虑：计算机图形学和计算机视觉。

计算机图形学和计算机视觉在侧重点方面有所不同。计算机图形学主要关注图像的生成和处理，其目的是通过算法和渲染、模拟等技术来生成出视觉效果，典型的应用包括电影制作、游戏开发等。而计算机视觉的侧重点在于从图像中获取和认知信息，其目的是从图像中捕捉到有用的特征信息，典型的应用包括目标检测[1]、人脸识别[2]等。

基于计算机图像学的风格迁移就是非真实感图形学（Non-photorealistic graphics，NPG）[3]，其又可以分成三类：基于笔触渲染思想的方法（Stroke-based Rendering,SBR）[4]、基于图像类比思想的方法（Image Analogy）[5]、基于图像滤波思想的方法（Image Filtering）[6]。基于计算机视觉的风格迁移主要是基于纹理合成的图像风格迁移，也就是纹理迁移。

2.1.1 基于笔触渲染思想方法

基于笔触渲染思想的图像风格迁移最早由Haeberli于1990年提出[7]，文章中提出了通过模拟绘画过程中笔触的移动来生成图像，为了实现高效的笔触渲

染，文中提出了一种基于 relaxation 的优化算法。该算法通过迭代优化笔触的位置和颜色，使得生成的图像逐渐趋向于期望的结果。

基于 relaxation 的优化算法是一种常用于解决约束优化问题的方法，它的核心思想是通过迭代地调整变量的取值，使得问题的约束条件得到满足。

该方法需要首先确定某种风格特征，不能随意的拓展成其他的风格。同时笔触渲染思想的方法需要大量的计算资源，而且风格较为单一，泛化能力差，不能随时切换成其他风格。

2.1.2 基于图像类比思想方法

基于图像类比思想方法的风格迁移最早由 Hertzmann 等人于 2001 年提出[4]，它通过将两个图像之间的相似性和差异性进行对比，利用两个图像之间的共同特征和结构信息，从而实现图像的风格迁移。首先需要一个具有所需目标风格的参考图像和一个需要进行风格迁移的目标图像，然后对两个图像进行特征提取，最后根据图像的相似度和差异性对目标图像进行调整，使得目标图像有更接近于参考图像的特征。

基于类比思想的图像风格迁移相比基于笔触渲染思想的图像风格迁移而言，能够实现多风格的迁移，但是基于类比思想的图像风格迁移需要成对的数据集，这种数据集在自然条件下较难采集，比如我需要采集春天和冬天两种风格下的人像图，采集出来的图片不可能人物姿态、动作、角度完全一样，同时基于类比思想的图像风格迁移可能会出现由于特征难以完全提取而出现图像失真的问题。

2.1.3 基于图像滤波思想方法

基于滤波处理思想的图像风格迁移是一种简单但有效的方法，它利用滤波器（也称为卷积核）[8]来调整目标图像的特征，从而使其具有参考图像的风格。使用基于图像滤波思想方法可以通过调整滤波器的参数可以模拟出不同风格种类，处理速度快、效果稳定，可以满足需要处理大量图像的场景，适合工业落地。但是基于滤波处理思想的图像风格迁移需要人为的调整滤波器的参数，才能得到想要得到最佳的渲染结果， 十分耗费精力。

2.1.4 基于纹理合成的方法

纹理合成是一种用于生成具有与给定样本纹理相似性的新纹理的技术。在计算机视觉领域，图像风格迁移被视作纹理合成的扩展。纹理合成的图像风格迁移是一种通过将两个图像之间的纹理信息进行匹配和合成，从而实现图像的风格迁移的方法。它可以将一幅图像的纹理特征转移到另一幅图像上，从而使得目标图像具有与参考图像相似的风格。

Efros 等人于 1999 年提出的模型使用了马尔可夫随机场非参数模型[9]，马尔可夫随机场[10]是一种用于描述随机变量之间相互依赖关系的概率图模型。在进行风格迁移的过程中，首先需要选择一个样本纹理，然后根据样本纹理和待合成图像的特征，构建一个能量函数，用于描述图像纹理的相似性和一致性。通常，能量函数包含两部分：数据项和正则项。数据项衡量合成图像与样本纹理之间的相似度，正则项则用来捕捉局部结构，接着将合成图像表示为一个马尔可夫随机场模型，其中每个像素被视为一个节点，节点之间的关系由能量函数来描述。最后使用迭代优化算法来最小化能量函数，从而得到最优的合成图像。

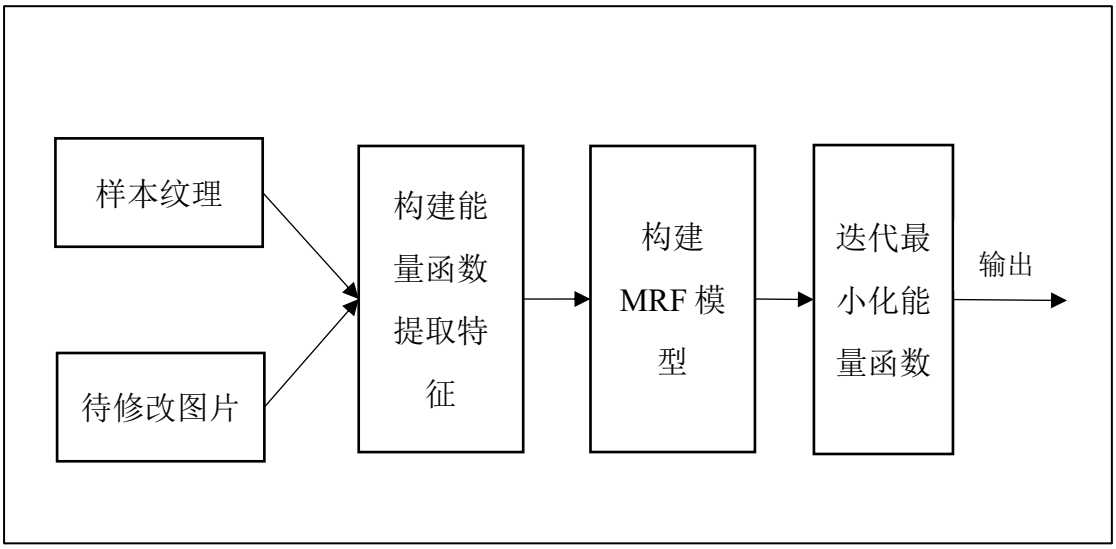

图 1 基于纹理合成的风格迁移流程图

我们可以将图像迁移看作图像纹理提取和新图像重建两个过程，随着深度学习[11]的发展，基于神经网络的图像风格迁移逐渐成了图像风格迁移领域研究的重点。

2.2 基于深度学习

2.2.1 基于在线图像优化的慢速图像迁移

这一类方法就是结合纹理提取和图像重建两个部分，Gatys 在 2015 年提出的”Texture Synthesis Using Convolutional Neural Networks”[12]就是利用了这一种方法，这方法的基本思路是以随机噪声为起始点，通过不断调整图像的每个像素值，使得最终得到的图像 x’ 的特征表示与我们设定的重建目标图像 x 的特征表示 Φ(x) 相似。换言之，迭代的目标是使得经过处理后的图像 x’ 的特征表示 Φ(x’) 尽可能接近于设定的目标特征表示 Φ(x)，即 Φ(x’)≈Φ(x)。如下图所示，输入是噪声底图，约束是 Style Loss 和 Content Loss 分别用来优化风格图像的风格和目标图像的内容。

在 Gatys 等人的设计中，该方法在卷积神经网络（VGG）[13]的基础上进行生成，文章中使用了 VGG-19 网络，在卷积神经网络中，每一层都包含一组非线性滤波器，这些滤波器的复杂性随着网络层次的增加而增加。当输入图像 x 经过卷积神经网络的每一层时，它会被编码成该层的滤波器响应。每个具有 N_i 个不同滤波器的图层生成大小为 M_i 的特征图，其中 M_i 是特征图的高度和宽度的乘积。第 L 层的响应可以被存储在矩阵 F_i 中，其中 F^l 是第 L 层中第 j 个位置处的第 i 个滤波器的激活。

$$L_{content}(\vec{p},\vec{x},l)=\frac{1}{2}\sum_{i,j}(F_{ij}^{l}-P_{ij}^{l})^2$$

为了展示在神经网络的不同层次中编码的图像信息，可以使用梯度下降算法对白噪声图像进行优化，以生成另一个图像，其特征响应与原始图像相匹配。设原始图像为 p，生成的图像为 x，它们在各自层级的特征表示为 P_l 和 F_l。然后，定义两个特征表示之间的平方误差损失。

为了捕捉输入图像风格的表示，使用了一个特征空间，该空间旨在捕获纹理信息。这个特征空间可以建立在网络的任何层的滤波器响应之上。它由不同滤波器响应之间的相关性组成，这些相关性的期望是在特征图的空间范围内获得的。这些特征相关性由 Gram 矩阵[14]给出。通过包括多层的特征相关性，获得了输入图像的静态、多尺度表示，其捕获了其纹理信息，但没有捕获全局排列。

可以通过构建一个与给定输入图像的风格表示相匹配的图像，将这些建立在网络不同层上的风格特征空间捕获的信息可视化。这是通过使用来自白噪声图像的梯度下降来最小化来自原始图像的 Gram 矩阵和要生成的图像的 Gram 矩阵的条目之间的均方距离来实现的。如图 2 所示，为了将艺术作品 a 的风格转移到照片 p 上，合成了一个新的图像，该图像同时匹配照片 p 的内容表示和艺术作品 a 的风格表示。

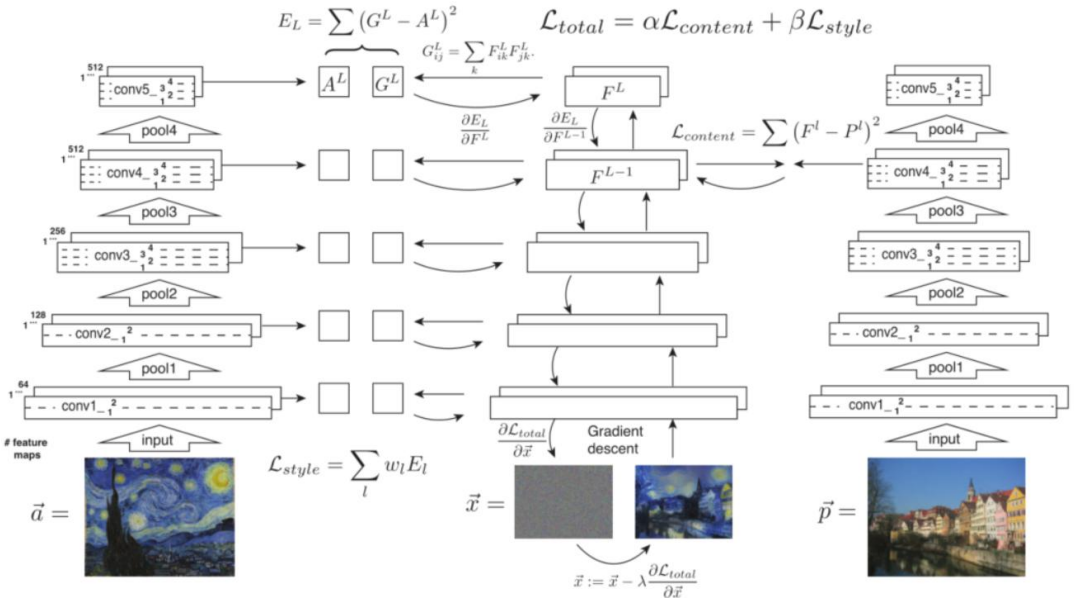

图 2 生成流程图

更容易理解的图像见图 3 所示。

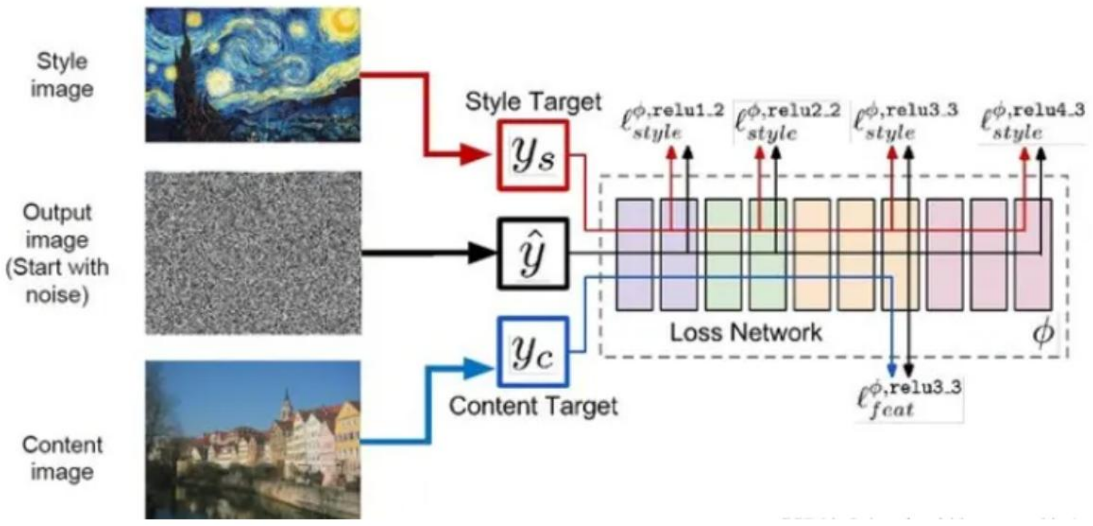

图 3 图像迁移流程图

可以看到，这个过程就是输入一张由随机噪声构成的底图，然后通过计算风格损失和内容损失，迭代更新底图，使其在风格上与样式图像相似，在内容上与原始照片相似。通常的训练过程是通过损失反向传播来更新网络参数，但在这里，我们使用一个预先训练好的 VGG16[15]作为基础模型，并锁定其参数，然后更新输入底图的像素值。更通俗的来说，我们需要将内容特征与风格特征融合在这个随机生成的噪声图中。我们很容易能够想到，这样生成是特别耗费时间经历的，每次迁移都需要对网络进行训练，但是这样的效果还是很好的。

2.2.2 基于离线图像优化的快速图像迁移

基于离线图像优化的快速图像迁移主要是为了解决上一节中提到的模型速度慢的缺点。在这个模型中，通过使用预先训练的前向网络，我们可以实现图像重建的快速方法，从而节省时间。这个前向网络经过训练后能够直接生成图像的重建结果，然后我们可以根据特定约束对结果进行优化。根据预训练的网络能够识别多少种风格作为分类标准，这种方法可以细分为单一模型单一风格（PSPM）[16]、单一模型多种风格（MSPM）[17]和单一模型任意风格（ASPM）[18]的快速图像风格迁移算法。

2.2.2.1 PSPM 算法

我们的主要思路是针对每种风格图像训练一个特定的前向网络[19]。这样，当进行测试时，我们只需将图像输入模型，经过一次前向计算即可得到输出结果，极大地节省了计算时间。这种方法根据纹理建模的不同方式，可分为两类：一是基于统计分布的参数化纹理建模，它着重于对图像纹理的数学统计特征进行建模，通过对图像局部区域的像素值分布、颜色直方图、梯度分布等进行统计分析，并利用概率模型来描述这些特征。其中较为经典的模型是由李飞飞教授提出的，模型如图 4 所示。

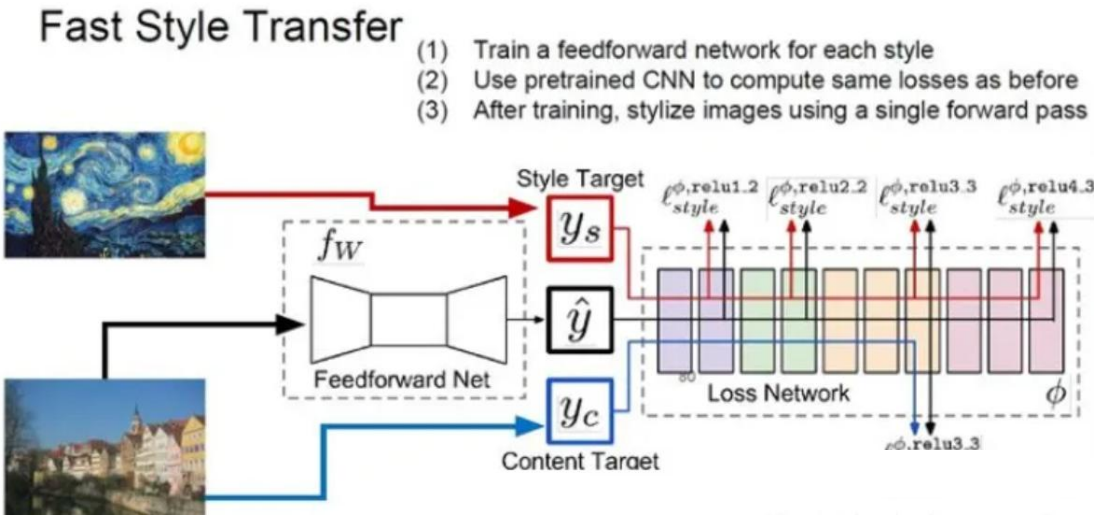

图 4 基于统计分布的参数化纹理建模

与 Gatys 提出的模型相比，新的模型的输入不再是随机噪声底图，而是目标图像。此外，新模型增加了一个类似于自动编码器的前向网络，用于拟合风格转移的过程。然而，与 Gatys 模型相似的是，新模型仍然使用基于另一个神经网络（通常是 VGG）提取的内容损失和风格损失，并将它们统称为 perceptual loss。在纹理建模方面，新模型也使用 Gram 矩阵来表达。

二是基于 MRF[10]的非参数化纹理建模，它更注重于描述图像纹理的空间结构和像素间的关系，它描述了图像中像素之间的空间关系，包括局部邻域的像素之间的相互作用。通过 MRF 模型，我们可以对图像的局部纹理结构进行建模，例如纹理的平滑度、连续性等。基于 MRF 的方法通常使用基于能量最小化的优化算法来实现图像重建或风格化迁移。

2.2.2.2 MSPM 算法

在单模型快速风格转移算法中，需要为每种风格训练一个模型。它只适合应用在风格种类较少的情况下，当需要处理成千上万种风格时，就会面临成千上万个模型的训练需求。这种情况下，这种方法在工业实践中将变得不可行。单模型多风格最早由 Dumoulin 等人发现[20]，其主要思想是利用不同风格网络之间的共享部分，并针对新的风格只调整其不同的部分，而保持共享部分不变。在训练好的一个风格化网络的基础上，通过使用 CIN 层进行仿射变换，可以将仿射变换的参数与每个风格进行绑定。对于每个新的风格，只需调整这些参数，而其余部分保持不变，就可以获得具有完全不同风格的结果。

MSPM 算法的核心在于把网络的一部分拿出来与每个特定的风格进行绑定，优点类似于大模型的微调，但是这种方法随着风格的增多会使得参数量也增多，因此使用这种方法虽然能够实现多种风格的切换，但是不能实现任意风格的切换。

2.2.2.3 ASPM 算法

ASPM 算法最早由 X.Huang 提出[17]，其受到 CIN 层的启发，提出了一个 AdamIN层，它将内容特征的均值和方差与风格特征的均值和方差对齐。它实现了与现有最快方法相媲美的速度，却不受对预定义风格集的限制。AdaIN 的输入是通过 VGG 网络提取的风格和内容特征。通过在大规模的风格和内容图像数据上进行训练，AdaIN 可以通过数据驱动的方式将图像中的内容 normalization 为成不同的风格。

2.2.3 基于生成对抗网络的风格迁移

2.2.3.1 生成对抗网络介绍

2017 年之后，生成对抗网络[21]渐渐的流行开来，生成对抗网络利用了统计学中的博弈论的思想，模型由一个判别器和一个生成器组成。生成对抗网络需要设置两个 loss，一个是生成器的 loss，loss 的目的是使得生成的图像能够被判别器判断为“真”，判别器使用另一个 loss，这个 loss 的目的是使得判别器能够辨别出正确的图像。

在图像迁移领域，生成对抗网络于 Gatys 模型[12]主要不同依然体现在 loss 的设计上。Gatys 模型的损失函数是通过 VGG 网络提取风格图片和目标图片的特征作为标签，然后提取生成图片的特征作为输出，并比较这两者的差异作为损失。相比之下，GAN 网络的损失函数设计更巧妙。GAN 网络由生成器和判别器两个网络组成，生成器负责将输入图片重建，然后将重建结果与真实数据集一起送入判别器进行评估。判别器的任务是区分生成器的输出结果是真实图片还是生成图片。这样生成器和判别器会形成一种动态的博弈，使得输出的图像风

格能够于数据集中的图像风格相似。

2.2.3.2 CycleGAN

我们在上一节提到的普通 GAN 模型可以用一句话进行简单的描述：输入噪声 X 通过生成器 G 得到输出 Y，Y 经过判别器 D 判断之后输出结果（是否为真实图像），从这一个过程中可以看到普通的 GAN 模型在生成的时候没有任何约束，完全是在“自由发挥”，为了模型能够更准确的输出想要风格的图片，需要给 GAN 模型加上约束。

CycleGAN 最早由 Jun-Yan Zhu 等人提出[9]，为了更好的生成图像，在一开始输入阶段便不再使用噪声作为输入，要想使用 CycleGAN 需要准备两个数据集，比如一个春天数据集，一个冬天数据集，输入时将春天数据集中的图片作为输入，判别时，将经过生成器重新构建的图片与冬天数据集的图片相对比，判断是否来自冬天数据集。这样生成器的 loss 就可以学到从春天数据集到冬天数据集的一个映射。CycleGAN 一个很大的优点就是，两个数据集不需要是成对的数据集，也就是说，春天数据集和冬天数据集只需要保证风格一直就行与主题内容无关（数据集中有明显的春天、冬天特征，比如春天枝芽萌发，冬天万物枯萎，数据集中可以有猫狗牛羊等任意生物，无需严格对等）。

到此，为了防止网络坍塌，即生成器发现一种图像可以骗过判别器就疯狂输出这种图像，判别器发现一个好的解就疯狂输出这个解的结果。CycleGAN 还需要对输出结果进行约束。

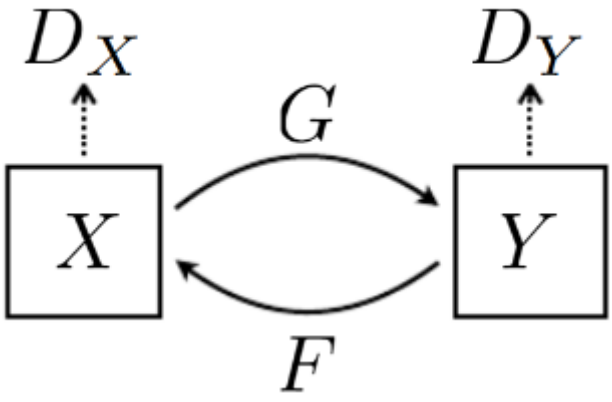

图 5 CycleGAN

如图 5 所示，模型需要对输出的结果进行一个反向变换，对比原始的输入

和反向变换后的输出的差异，依然那春天数据集和冬天数据集来举例子，输入春天数据集的图片，判别器D_Y是用来判别输出的图片是否像冬天数据集，而判别器D_X是用来判别输出的图片和输入图片（春天数据集中的图片）是否有关系。

CycleGAN 由四部分组成，两个判别器两个生成器和 4 个 loss，通过内容约束实现效果更好的图像风格迁移。

2.2.3.3 StarGAN

从上一小节的叙述中我们可以发现 CycleGAN 实际上属于一种单模型单风格的迁移，我们如果想要实现多个风格的相互转换就需要很多的判别器和生成器。为了实现单模型多风格的生成 StarGAN 就被提出来了。

StarGAN[22]最早由 Yunjey Choi 等人提出，可以实现图像的多领域之间切换，

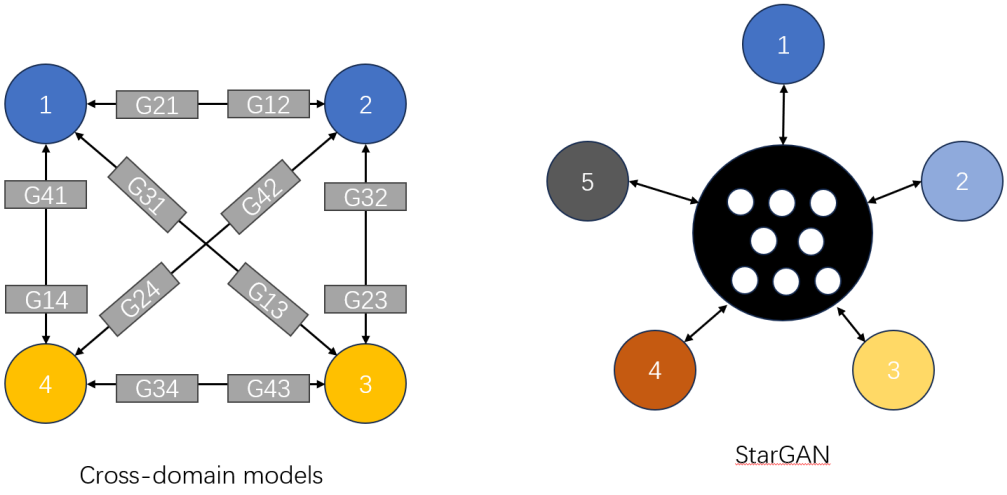

图 6 StartGAN 与其他 GAN 的区别

如图 6 所示，左边是普通的 GAN 模型，模型有多个生成器需要在不同领域转移期间进行特征提取，而 StarGAN 只需要一个生成器。

如图 7 中 a 图所示，StartGAN 同样由生成器和判别器两部分，鉴别器 D 需要学习辨别真实图像和生成图像，同时输出图像所属分类。为了实现图像域的准确转换，StarGAN 在判别器的顶部引入了一个复杂的辅助分类器。在同时优化生成器（G）和判别器（D）的过程中，它添加了一个域分类损失。判别器的训练目标是最小化真实图像的域分类损失，以学习真实图像的正确分类；而生成器的训练目标是最小化生成图像的域分类损失，以确保生成的图像能够被正

确地分类为目标域。

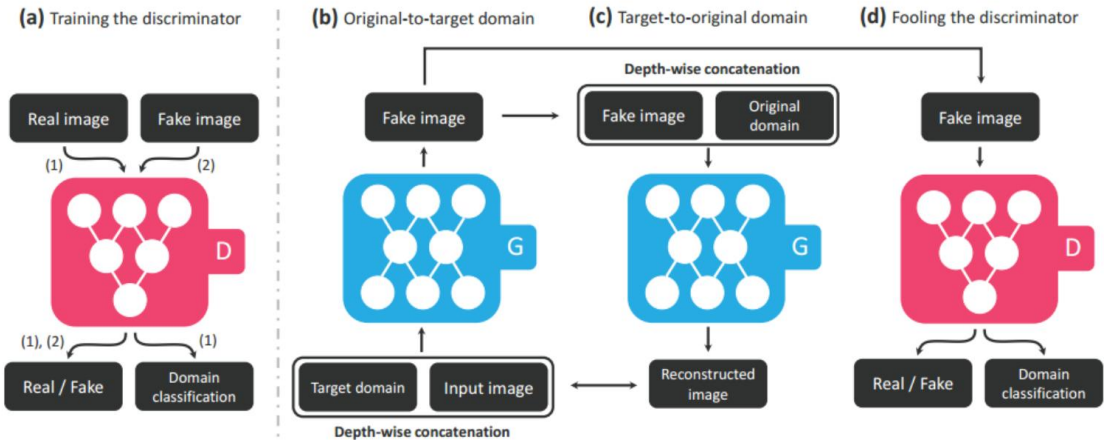

图 7 StarGAN 流程

为了保留输入图像上与域转换无关的内容，作者参考了 CycleGAN 的处理方式，使用两次生成器，一次将输入图像转换成目标图像，然后再使用一次生成器将目标图像重构成输入图像，比较原始图像与重构图像的 L1 正则化距离。

因此，StarGAN 的损失函数有 5 个，判别器 2 个，生成对抗损失和分类损失，生成器有 3 个损失函数，对抗损失，分类损失和重构损失，作者将分类损失权重设置为 1，重构损失权重设置为 10。

2.2.4 基于自然语言语义信息（textual）指导的风格迁移

2.2.4.1 CLIP 模型

CLIP[23]全称 Constrastive Language-Image Pre-training，是 OpenAI 推出的采用对比学习的文本-图像预训练模型。CLIP 惊艳之处在于架构非常简洁且效果好到难以置信，在 zero-shot 文本-图像检索[24]，zero-shot 图像分类[25]，文本-图像生成任务 guidance[26]，open-domain 检测分割[1]等任务上均有非常惊艳的表现。

原始的 CLIP：基于对比学习[27]，在大量图-文对数据上进行训练，让图像特征和文本特征在同一个向量空间中对齐；这个空间包含了图像和文本域的相关信息，因此我们可以根据自然语言得到其在图像层面的特征，进一步控制风格迁移的过程；

发展的 CLIP：预训练数据越来越多，特征对齐完成度很高，文本信息能够和多层次的图像特征对齐，在控制风格迁移时效果越来越好；

2.2.4.2 语义信息（textual）指导风迁移

基本思想：对于原域和目标域有准确的文本描述，网络输出图像和原域图像在 CLIP 向量空间中的距离应当与文本的距离一样，这个过程用对比损失约束；

语义信息的约束关键在输出端构建损失，而网络是不受限制的，可以是最初始的 VGG[28]、CycleGAN，也可以是 StyleGAN[29]；

CLIPstyler: Image Style Transfer with a Single Text Condition [30]具有单一文本条件的图像风格迁移中指出：

现有的神经风格迁移方法需要参考风格图像将风格图像的纹理信息迁移到内容图像。然而，在许多实际情况下，用户可能没有参考的风格图像，但仍然有兴趣通过想象来传递风格。为了处理此类应用需求，该文提出了一个新框架，该框架可以在“没有”风格图像，只有所需风格的文本描述的情况下实现风格迁移使用预训练文本-图像嵌入模型 CLIP。用广泛的实验结果证实了利用反映语义查询文本的真实纹理进行的成功的图像风格迁移。

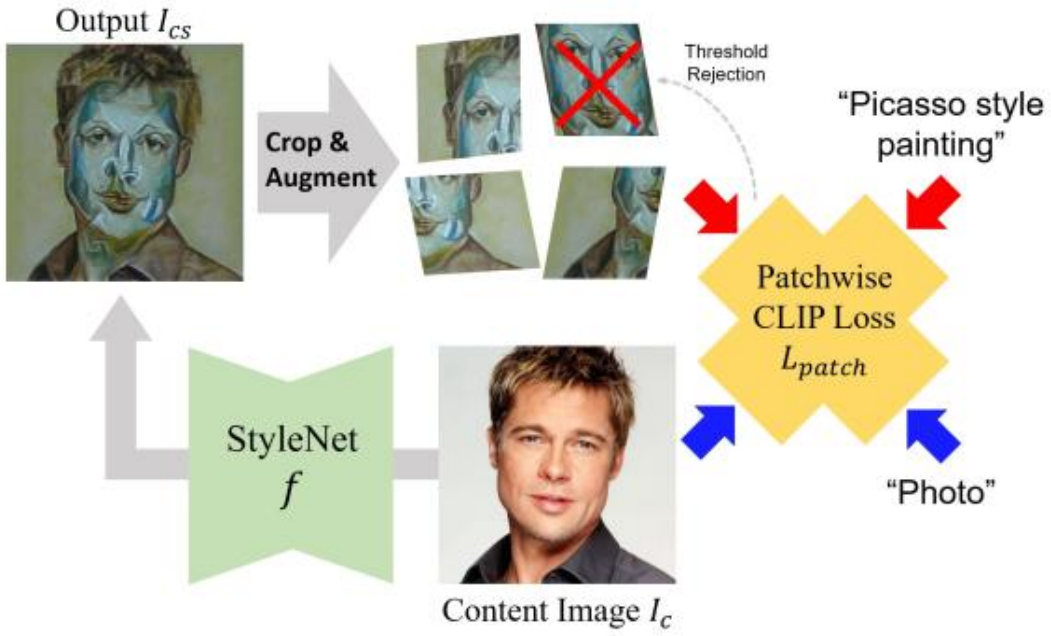

图 8 CLIPstyler 模型图

图 8 是该文的基本框架，目的是通过预先训练的 CLIP 模型，将目标文本的语义风格迁移到内容图像上。与现有方法的不同之处在于，该文没有风格图像作为参考。

2.2.5 基于扩散模型的风格迁移

扩散模型[31]是生成模型的一种，可以很好的解决 GAN 模型生成效果不稳定的缺点。扩散模型最早是在 2015 年提出的，可以被视为去噪自编码器，用来消去高斯噪声。真正大规模应用是在 2020 年 DDPM 将扩散模型用于图像生成。

2.2.5.1 扩散模型核心介绍

扩散模型的核心是迭代，如图 9 所示，扩散模型与其他生成模型具有明显的不同，首先扩散模型各种层之间是同维度的，其次扩散模型存在前向和反向两个过程。

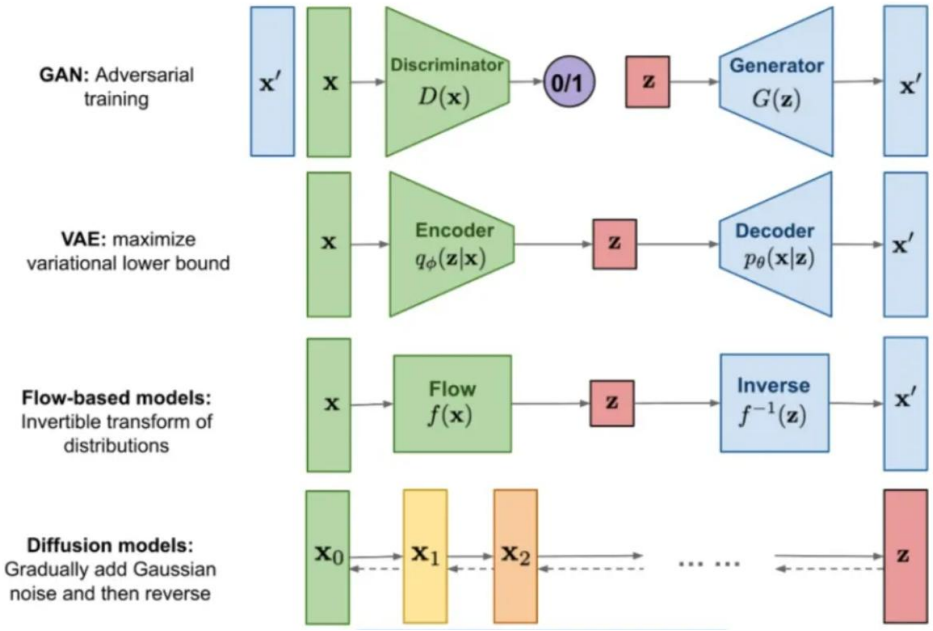

图 9 几种生成模型结构图

那么扩散模型具体的工作过程就是前向过程和后向过程两个过程，前向过程负责给图片加高斯噪声，后向过程负责给图片减去噪声。

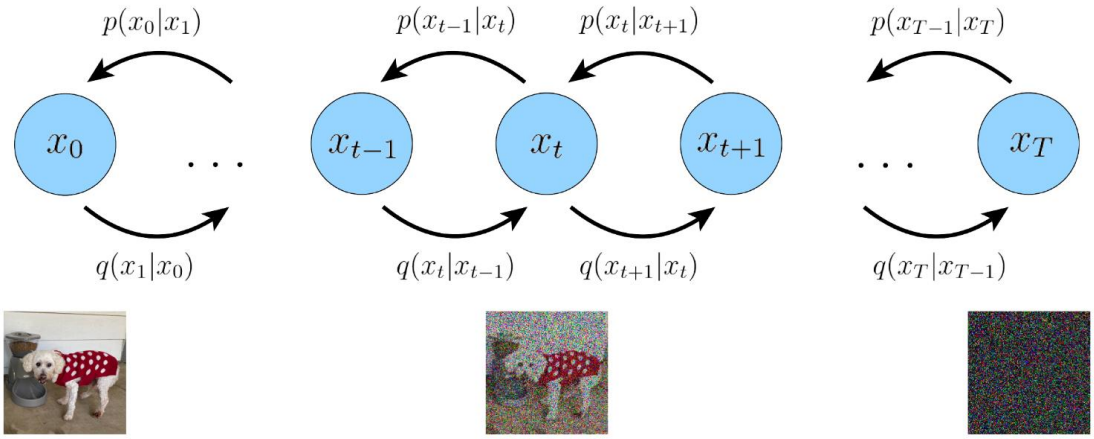

图 10 扩散模型原理图

如图 10 所示，扩散模型的前向过程简单来说就是不断向图片里面加入噪声，直到图片变的面目全非，也就是从x_0到x_T的过程，我们可以从图片中观察到，x_t时刻的状态只与x_{t-1}时刻有关，说明扩散模型是符合马尔科夫定理的。

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}z_1$$

论文中给出了前向过程的公式如上述公式所示，α_t是一个权重，它会随着 t 的增大而减小，代表着随着 t 的增大，噪声所占越多，需要添加的噪声也越多。我们如果要加入 1000 次噪声只需要根据公式（1）进行递推就好了。

我们知道了前向过程想要求反向过程，可以发现前向过程和反向过程只是将括号里的两个变量互换了一下位置，因此，可以由贝叶斯公式进行反向过程的推导。推导结果如下述公式所示。

$$\tilde{\boldsymbol{\mu}}_{t-1}(\mathbf{x}_{t-1}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\mathbf{z}_t)$$

现在发现想要求解反向过程需要知道$\mathbf{z}_t$，$\mathbf{z}_t$在前向过程符合高斯分布，但是在后向过程就不一定符合高斯分布了，因此我们需要学习通过x_t来预测$\mathbf{z}_t$，既然要使用机器学习。使用机器学习就需要有标签，这里的标签就是前向传播过程中，每次迭代从高斯分布采取到的噪声。训练模型的时候，模型根据x_t来预测$\mathbf{z}_t$，然后通过$\mathbf{z}_t$根据上述公式来计算 t-1 时刻的分布均值。

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

图 11 训练流程伪代码

在模型训练好之后，必须从 T 时刻（全是噪声）向前逐步生成图片，论文中的流程如图 12 所示，先生成一个标准高斯分布噪声，在每个时间步中，会将上一步生成的图像x_t输入模型中预测噪声，然后预测x_{t-1}，直到第一个时间步。

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

图 12 预测流程伪代码

2.2.5.2 基于扩散模型的风格迁移

扩散模型可以适用于一切图像生成领域，同样，在风格迁移领域应用风格迁移可以使得模型仅需一小组图像就能够稳定的生成特定风格的图像。

在“Diffusion in Style”[32]一文中，作者将图像的风格看作特征分布，使用扩散模型实现了自定义风格图像的生成。文章中模型的关键在于使用初始潜在张量去降噪，只需要使用一小组目标风格的图像就可以获得初始潜在张量的风

格特定分布。

具体而言，作者没有改变前向过程的类型，仅仅通过改变噪声分布的位置和协方差，使得模型可以不需要重新进行大量的训练。如图 13 所示，传统的扩散模型中的噪声是符合标准正态分布，在利用扩散模型进行风格迁移的时候需要将标准的正态分布噪声改成适用于特定样式的噪声分布，然后使用一小组图像进行微调训练。这方法主要包括两个关键步骤：首先，通过一组目标图像来计算适应所需风格的噪声分布；然后，将这个适应风格的噪声分布应用到扩散模型的 UNet 网络中，以生成符合所需风格的图像。

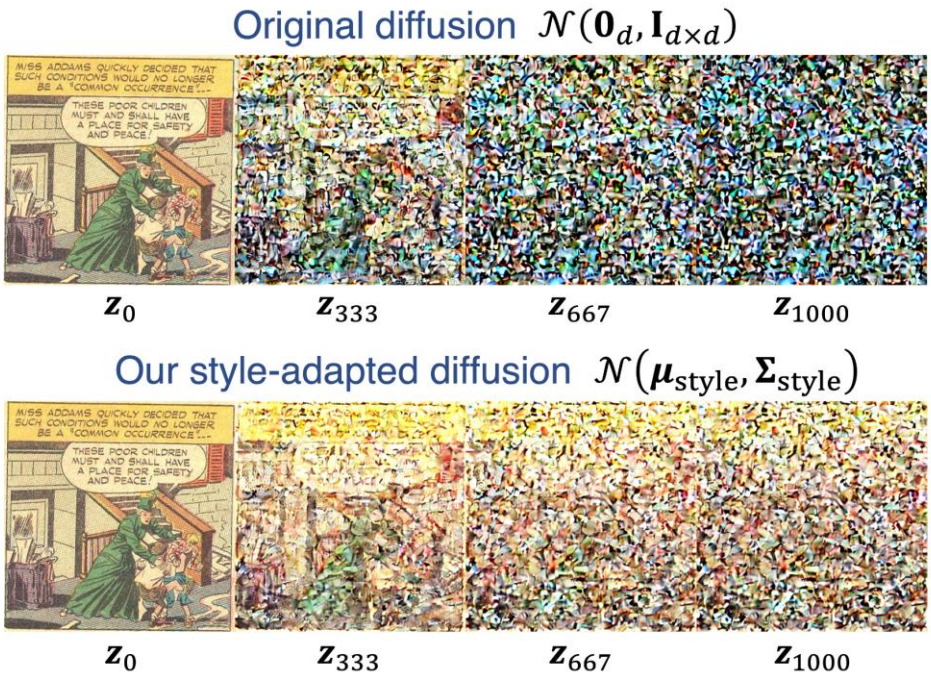

图 13 扩散模型改进

三、实验设置

3.1 风格迁移任务的挑战

经过上述论文综述的调查，本文发现风格迁移任务在如今面临的挑战：

1. 风格与内容的平衡：在风格迁移过程中，如何在保持源内容图像的主体结构和特征的同时，恰当地融入目标风格，是一个核心挑战。

2. 用户可控性：为用户提供一定程度的控制权，如调整风格强度、选择特定风格区域应用等，可以增加风格迁移的实用性和艺术创作的灵活性，但这也增加了算法设计的复杂度。

3. 训练数据限制：获取高质量、多样化的训练数据集是一大挑战，特别是对于某些独特或小众的艺术风格，数据稀缺可能限制了模型的学习能力和泛化性能。

针对如上三个问题，本文研究与运用了 CAP-VSTNet 模型[33]、DEADiff 模型[34]与 LoRA 微调[35]，期望能够运用最新的技术在一定程度上对上述问题进行改善。以此为思路，本文设计了三个模型的实验。

3.2 CAP-VSTNet 模型

CAP-VSTNet 模型是一个在 2023 年提出的风格迁移网络，它在处理风格迁移任务时表现出了优秀的性能。这个网络包括一个新的可逆残差网络和一个无偏线性变换模块，用于多功能风格转移。这个可逆残差网络不仅能保留内容亲和度（包括像素亲和度和特征亲和度），而且不会像传统的可逆网络那样引入冗余信息，从而实现更好的风格化。借助 Matting Laplacian 训练损失，可以解决由线性变换导致的像素亲和力损失问题，所提出的框架适用于多种风格迁移。其有效的解决了上文所说的风格与内容的平衡问题。

CAP-VSTNet 的主要目标是解决内容相似度损失问题，这是导致逼真和视频风格迁移中出现伪影的主要问题。根据相关研究，CAP-VSTNet 在多功能风格转移上表现出了有效性，并且可以产生较好的定性和定量结果。这意味着 CAP-VSTNet 能够在保留内容相似性的同时，实现高质量的风格迁移。

如图 14 所示，给定内容图像和风格图像，该框架首先通过网络的前向推理

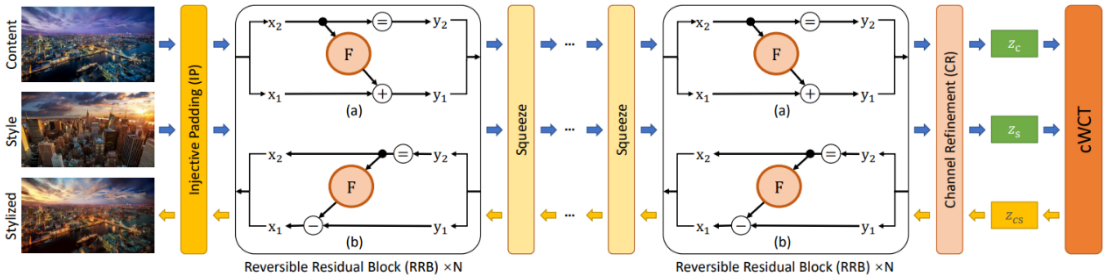

图 14 CAP-VSTNet 模型框架结构

将输入内容/风格图像映射到潜在空间，然后 injective padding 模块通过沿通道维度的零填充来增加输入维度。前向推理是通过级联的可逆残差块(Reversible Residual Network)和空间压缩模块(Squeeze)执行的。之后，使用通道细化(Channel Refinement)去除内容/风格图像特征中的通道冗余信息，以实现更有效的风格转换。然后使用线性变换模块 (cWCT) 来传输内容表示以匹配样式表示的统计信息。最后，通过后向推理将风格化表示反向映射风格化图像。

3.2.1 CAP-VSTNet 模型评估

真实感测试和艺术感测试是用来评估风格迁移模型 CAP-VSTNet 的性能的两种测试方法。这些测试方法帮助研究者们理解模型在进行风格迁移时，如何平衡内容的真实性与艺术风格的表达。

真实感测试：主要关注模型是否能在风格迁移后保持原始内容的真实性。，评估这些生成图像与原始内容图像在视觉上的相似度，以及它们是否保留了原始图像的关键特征。

艺术感测试：侧重于评估风格迁移后图像的艺术风格表达。这通常涉及到对风格图像和生成图像之间的风格相似度进行评估。这种测试的目的是确保风格迁移后的图像不仅保留了原始内容，同时也成功地吸收了目标风格图像的艺术特征。

实现这些评价的方法通常包括定量和定性的分析。定量分析可能涉及计算图像之间的相似度分数，例如使用结构相似性指数（SSIM）[36]或峰值信噪比（PSNR）[37]。定性分析则可能包括由人类观察者进行的视觉评估，他们会根据图像的真实感和艺术感给出评分。

3.2.2 实验设置

数据集准备：

MS_COCO 数据集[38]：MS_COCO 数据集是一个大型的现实生活照片数据集，包含多种多样的图像，适用于风格迁移模型的训练。

Link 数据集[39]: 由 6656 张真实的风景照片组成，涵盖了三种独特的动漫风格：Hayao，Shinkai，和 Paprika。每种风格的动漫图像都是通过从相应电影的

视频帧中随机裁剪得到的。

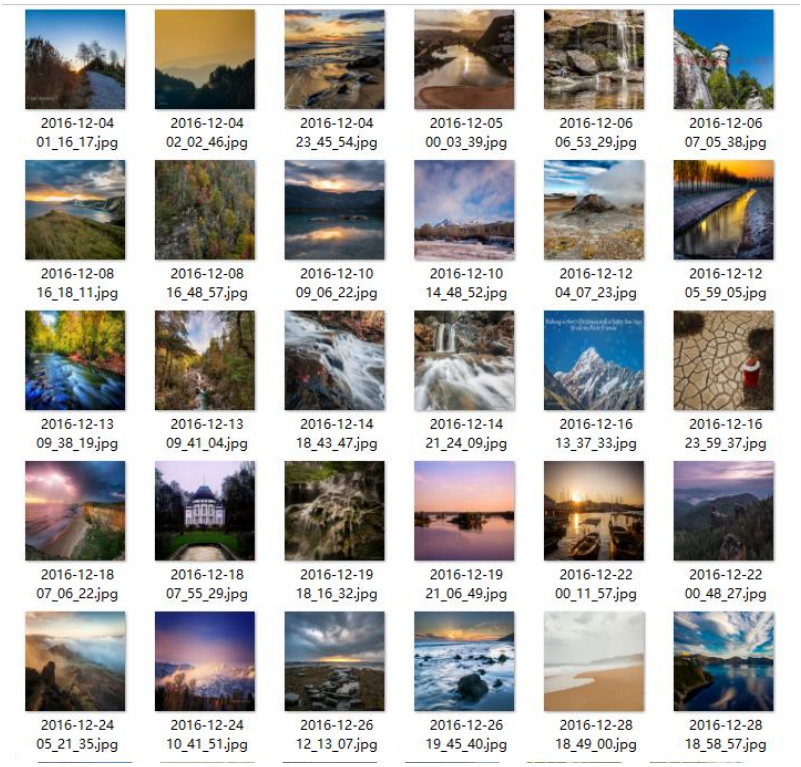

图 15 MS_COCO 数据集

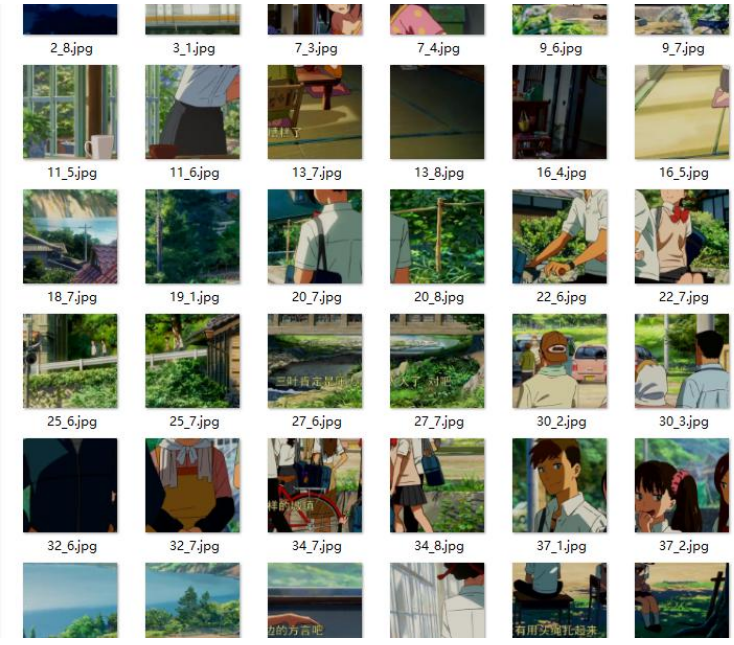

图 16 Link 数据集

训练过程：

模型选择：选用了预训练的 Pre-Trained VGG19 模型

训练配置参数修改：根据实验需求调整训练脚本中的配置参数，例如 batch size、学习率等，整体迭代了 170000 轮。

开始训练：修改好参数后，可以通过运行训练脚本或使用命令行来启动训练过程。

测试过程：

单帧图片测试和视频测试：训练完成后，可以对单帧图片或视频进行风格迁移测试，以评估模型的性能。

测试配置参数修改：在测试之前，需要修改测试脚本中的配置参数，指定模型权重文件、内容图像和风格图像等。

3.3 DEADiff 模型

DEADiff 模型[34]是一种用于风格迁移任务的扩散模型，它在 CVPR2024 会议上被介绍。该模型的主要特点是它能够解耦参考图像的风格和语义表示，从而在不损失文本条件可控性的情况下，有效地将参考风格转移到合成图像上。其主要解决上文中的用户可控性问题。

DEADiff 模型采用了以下两种主要策略：

解耦风格和语义的机制：DEADiff 通过 Q-Formers 提取解耦的特征表示，这些特征表示受不同的文本描述指导。然后将它们注入到互斥的交叉注意力层的子集中，以更好地解耦风格和语义。

非重构性学习方法：与传统的扩散模型不同，DEADiff 使用成对图像进行训练，而不是相同的目标图像。在这种情况下，参考图像和真值图像具有相同的风格或语义。

这些策略使得 DEADiff 在保持文本到图像模型中固有的文本可控性的同时，也能够实现与参考图像的风格相似性，从而在视觉风格化结果上获得了最佳平衡。DEADiff 的效率和风格转移能力都得到了提升，与基于优化的方法相比，它更加高效，同时保持出色的风格转移能力。与传统的基于编码器的方法相比，DEADiff 可以有效地保持文本的控制能力

3.3.1 准备工作

数据集准备：为了训练 DEADiff 模型，需要准备一组包含风格和内容信息的图像数据集。这些数据集通常由成对的图像组成，其中一张图像定义了风格，

另一张图像定义了内容。

解耦风格和语义：

a. 特征提取机制：

论文中介绍了一种双重解耦表示提取机制（DDRE），它利用 Q-Former 从参考图像中获取风格和语义表示。Q-Former 通过“风格”和“内容”条件进行指导，选择性地提取与给定指令对齐的特征。

受BLIP-Diffusion[40]的启发，后者通过具有不同背景的合成图像对学习主体表示，以避免琐碎解。整合了两个辅助任务，利用 Q-Formers 作为表示过滤器，嵌套在非重构范式中。这使能够隐式地识别图像中风格和内容的解耦表示。

一方面，采样一对不同的图像，两者保持相同的风格，但分别作为 Stable Diffusion（SD）生成过程的参考和目标，如图 17(a)对所示。

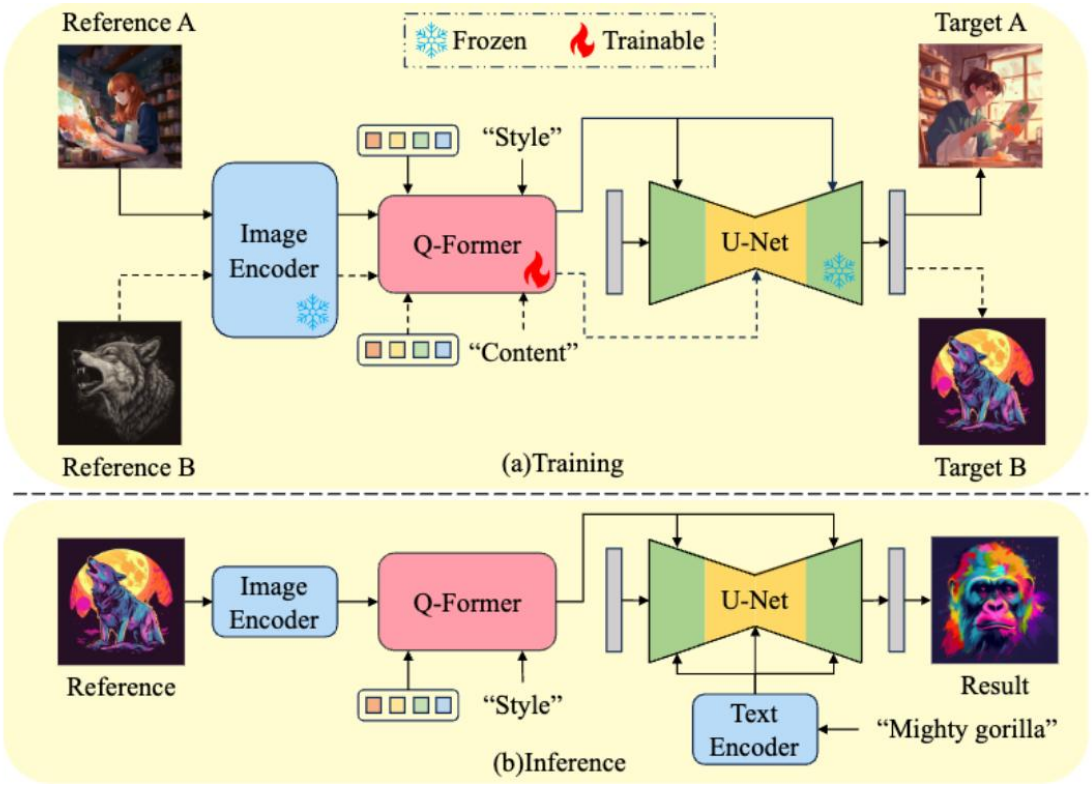

图 17 双重解耦表示提取机制示意图

b. 解耦条件机制：

不同交叉注意力层在去噪 U-Net 中主导合成图像的不同属性的启发，引入了一种创新的解耦条件机制（DCM）。本质上，DCM 采用了一种策略，将具有较低空间分辨率的粗层置于语义条件下，而具有较高空间

分辨率的细层则置于风格条件下。

如上图 17(a)所示，仅将具有“风格”条件的 Q-Former 的输出 query 注入到细层中，这些细层响应于局部区域特征，而不是全局语义。这种结构性的调整促使 Q-Former 在输入“风格”条件时提取更多的风格导向特征，例如图像的笔触、纹理和颜色，同时减弱了其对全局语义的关注。因此，该策略使得风格和语义特征的解耦更加有效。

同时，为了使去噪 U-Net 支持图像特征作为条件，设计了一个联合文本-图像交叉注意力层，如图 18 所示。

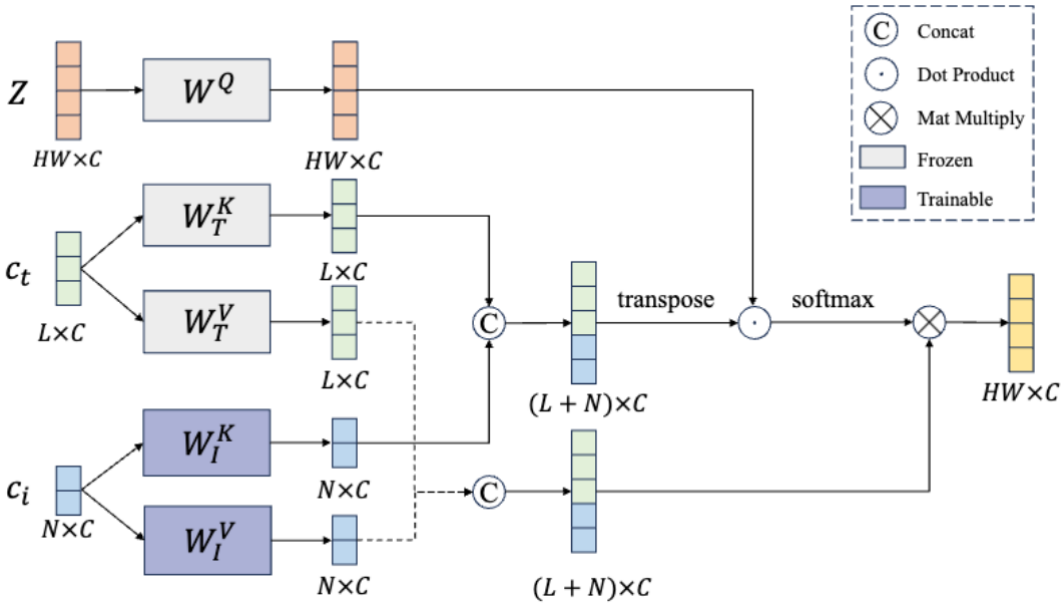

图 18 联合文本-图像交叉注意力层

与 IP-Adapter 类似，包含了两个可训练的线性投影层，来处理图像特征，与文本特征一起，还包括了冻结的线性投影层。然而，不是独立地对图像和文本特征执行交叉注意力，而是分别从文本和图像特征中连接键和值矩阵，随后使用 U-Net query 特征 Z 进行单个交叉注意力操作。

非重构训练范式：使用具有相同风格或相同语义的图像对进行训练，以分别优化风格和语义 Q-Former，避免重构任务对模型的偏差影响。

3.3.2 实验设置

模型训练：DEADiff 模型使用成对图像进行训练，而不是相同的目标图像。这意味着参考图像和真值图像具有相同的风格或语义。

该部分实验采用预训练模型，其训练细节为采用了 Stable Diffusion v1.5 作为基础文本到图像模型。此模型包括 16 个交叉注意力层，其中 4-8 层定义为粗层，用于注入图像内容表示，其余层定义为细层，用于注入图像风格表示。图像编码器采用 CLIP 的 ViT-L/14，Q-Former 的可学习查询 token 数量为 16。两个 Q-Formers 分别提取语义和风格表示，并使用 BLIP-Diffusion 的预训练模型进行初始化。

3.4 LoRA 微调

LoRA 微调[35]（Low-Rank Adaptation，低秩适应）是一种微调技术，它在深度学习模型中引入了低秩矩阵来调整预训练模型的权重，而不是直接修改原始权重。这种方法在保持预训练模型结构的同时，通过增加少量可训练参数来适应新任务，从而减少了微调过程中的计算复杂性和资源消耗。其一定程度解决了前文说的训练数据限制问题。

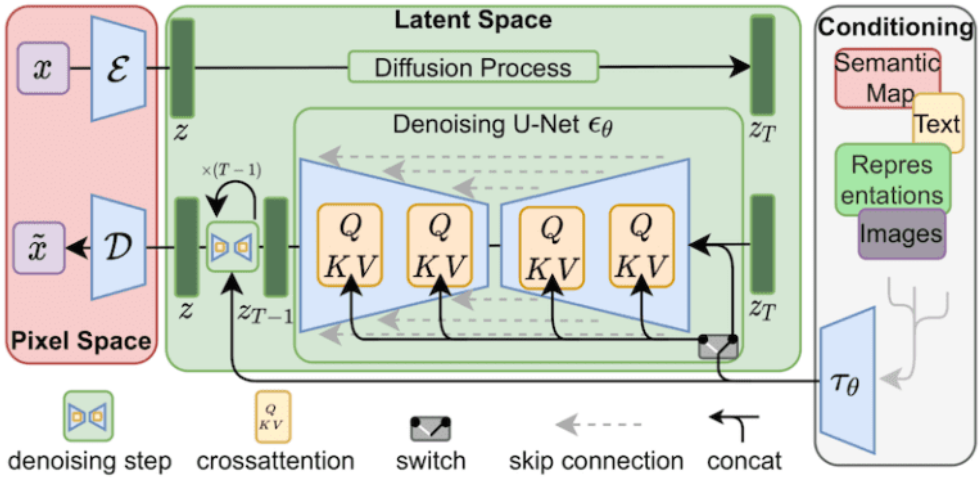

图 19 LoRA 微调注意力层

在风格迁移任务中，LoRA 微调可以用来调整模型以更好地适应特定的风格特征。具体来说，LoRA 通过对模型中的注意力层（如 Transformer 模型[40]中的自注意力层）进行低秩分解，只更新一个小的参数子集，而不是整个权重矩阵。这样做的好处是可以显著减少需要训练的参数数量，同时保持模型的性能。

LoRA 微调的核心思想是在原有的权重矩阵上添加一个低秩矩阵，这个低秩矩阵由两个较小的矩阵的乘积构成。这样，原始的权重矩阵保持不变，而通过训练这两个小矩阵来实现对模型的微调。这种方法在多个领域，包括自然语言

处理和计算机视觉中都显示出了良好的效果。

3.4.1 模型选择

本文基于 tAnimeV4Pruned_v40 模型使用 LoRA 微调。tAnimeV4Pruned_v40 模型是一个基于 Stable Diffusion 的风格迁移模型，专门用于生成具有日漫风格的图像。该模型经过剪枝处理，使得模型变得更小，更适合在资源有限的环境中使用。它在细节上进行了优化，比如增强了画面场景的能力和图像的稳定性，改善了肢体问题，并且在语义识别和脸部特效上也有所加强。

3.4.2 实验设置

LoRA 微调设置：在模型的特定层中引入低秩矩阵，并进行初始化。这些低秩矩阵将用于微调过程中的参数更新。

训练数据：本文选用了自己找寻的属于中世纪、暗黑风、高清的游戏风格

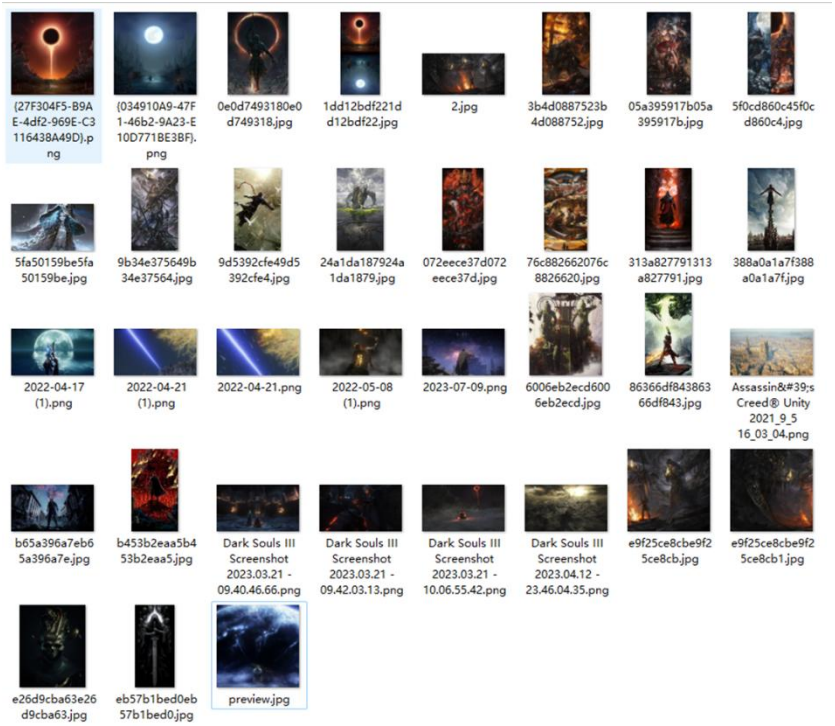

图 20 LoRA **模型微调训练数据集**

图片（如图 20）

微调工具选用：选用集成很好的模型训练器 Cybertron Furnace。

Cybertron Furnace（赛博丹炉）是一个专注于 Windows 系统的 Gradio 图形

用户界面（GUI），用于 Kohya 的 Stable Diffusion 模型训练器。它允许用户设置训练参数，并生成及运行所需的命令行指令来训练模型。这个工具集成了必要的依赖项和配置文件、预训练脚本，支持人物、二次元、画风等 LoRA 模型的训练，旨在简化用户训练 LoRA 模型的流程。它还支持图片预处理、图片标签编辑，以及查看训练进度等功能。使用该训练器极大的简化了本文的 LoRA 模型的训练。

模型保存和部署：微调完成后，将训练好的低秩矩阵与原始模型的权重合并，保存为新的模型版本，以便于后续的使用和分析。

四、实验结果与分析

4.1 CAP-VSTNet 模型

以图 21 中的古风、偏水墨风的风格为例：

图 21 古风、偏水墨风图片

将图 22 输入本文训练的 CAP-VSTNet 模型得到效果如图 23：

图 22 云南大小图书馆（侧面）

图 23 CAP-VSTNet 模型效果图

本文引入 CAP-VSTNet 模型是为了解决风格与内容平衡的问题，即过度的风格应用可能会淹没原始内容，而风格不足则达不到预期的艺术效果。由图 23 可以看出，经过模型得到的图片即有了图 21 中的水墨风（针对建筑古风的特点不够明显）又保持了原图 22 的内容，并没有对原图的内容增加或修改。模型的效果符合本文的预期。

4.2 DEADiff 模型

任然以图 21 的风格为例，本文设置 prompt 为：there is a clock tower in the forest(在森林中的钟塔)，得到图 24 的效果：

图 24 DEADiff 对应 prompt 效果图

本文也对图 24 的效果感到惊喜，其对水墨风、古风的诠释很优秀如其绘制的古风的钟楼、钟楼与树木将的布局。

引入 DEADiff 模型，是为了解决本文发现的风格迁移中的用户可控性，于是本文设置了下述生成：

图 25 云南大学图书馆侧面照

图 26 用户可控性尝试的 prompt 设置

Chinese palace, Chinese architecture, high quality architectural art

The front of a building

the starry night, Vincent van Gogh, 1884, oil on

本文以 25 图为原始图片，加入了图 26 中所设置的 prompt：

在该 prompt 中，黑字主要设置及强调图片的风格为中国古风、唐代建筑、高质量高精度；用户可控性的体现为 prompt 中的生成建筑正面（红字）、生成天空为梵高的星空夜（蓝色）。最终得到效果图 27

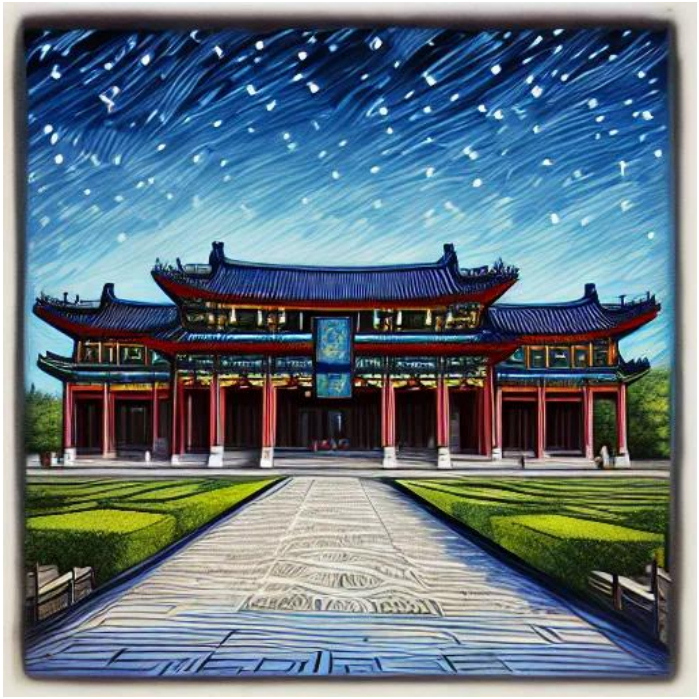

图 27 用户可控性生成测试效果图

由图 27 可以发现，模型不仅实现了生成建筑的正面照、天空为梵高的星空夜油画风格的要求，还能生成十分精细的文案。本文对此效果十分震撼，但与此同时，本文发现其对原始图片的内容保留的效果不仅如意，可能是本文的 prompt 设置的不够正确，又或许是模型本身的缺陷。

4.3 LoRA 微调

任然以图 25 为原始图片，本文训练的 LoRA 微调模型效果如图 28

图 28 LoRA 微调模型效果图

观察图 28，本文发现其确实体现了中世纪时的建筑风格，但是对原始图像的内容也一定程度进行了修改。但是本文引入 LoRA 微调模型，是为了解决训练数据限制问题。而本文只使用了少量的高清图片为数据集，便能训练出此效果的模型。在训练数据限制的情况下，无疑是一个优秀的方向。

五、总结

目前，风格迁移领域已经取得较多的成果，特别是深度学习技术发展起来之后，风格迁移的效果变得越来越好也越来越接近现实。

本文针对经过前期调研发现的风格迁移中的三个难题：风格与内容的平衡难题、实现用户可控性、训练数据限制，分别使用CAP-VSTNet模型、DEADiff模型、LoRA微调模型尝试解决。这三个模型在一定程度上都达到了其解决对应问题的能力。本文也为模型优秀的效果所震撼。

但与此同时本文在模型训练中也发现 CAP-VSTNet 模型在处理复杂纹理时难度非常高；LoRA 存在一些较大的瑕疵，包括参数的调优，TAG 的生成与最后 Prompt 的设置。本文在本次任务中还有许多不足，如创新性有待提升，目前基本上都是在使用别人的模型。

参考文献

[1] Koopman B O. The theory of search. II. Target detection [J]. Operations research, 1956, 4(5): 503-31.

[2] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 815-23.

[3] Strothotte T, Schlechtweg S. Non-photorealistic computer graphics: modeling, rendering, and animation [M]. Morgan Kaufmann, 2002.

[4] Hertzmann A. A survey of stroke-based rendering [C]. 2003. Institute of Electrical and Electronics Engineers.

[5] Hertzmann A, Jacobs C E, Oliver N, et al. Image analogies [M]. Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 557-70.

[6] Chandel R, Gupta G. Image filtering algorithms and techniques: A review [J]. International Journal of Advanced Research in Computer Science and Software Engineering, 2013, 3(10).

[7] Redfield A G. On the theory of relaxation processes [J]. IBM Journal of Research and Development, 1957, 1(1): 19-31.

[8] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2414-23.

[9] Zhu J-Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]. Proceedings of the IEEE international conference on computer vision, 2017: 2223-32.

[10] Cross G R, Jain A K. Markov random field texture models [J]. IEEE Transactions on pattern analysis and machine intelligence, 1983, (1): 25-39.

[11] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521(7553): 436-44.

[12] Gatys L, Ecker A S, Bethge M. Texture synthesis using convolutional neural networks [J]. Advances in neural information processing systems, 2015, 28.

[13] Wen L, Li X, Li X, et al. A new transfer learning based on VGG-19 network for fault diagnosis [C]. 2019 IEEE 23rd international conference on computer supported cooperative work in design (CSCWD), 2019: 205-9. IEEE.

[14] Sreeram V, Agathoklis P. On the properties of Gram matrix [J]. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 1994, 41(3): 234-7.

[15] Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification [C]. 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON), 2021: 96-9. IEEE.

[16] Pataky T C, Goulermas J Y. Pedobarographic statistical parametric mapping (pSPM): a pixel-level approach to foot pressure image analysis [J]. Journal of biomechanics, 2008, 41(10): 2136-43.

[17] Wang Y, Si Y, Huang B, et al. Survey on the theoretical research and engineering applications of multivariate statistics process monitoring algorithms: 2008–2017 [J]. The Canadian Journal of Chemical Engineering, 2018, 96(10): 2073-85.

[18] Ghiasi G, Lee H, Kudlur M, et al. Exploring the structure of a real-time, arbitrary neural artistic stylization network [J]. arXiv preprint arXiv:170506830, 2017.

[19] Fine T L. Feedforward neural network methodology [M]. Springer Science & Business Media, 1999.

[20] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans [J]. Advances in neural information processing systems, 2017, 30.

[21] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-44.

[22] Choi Y, Uh Y, Yoo J, et al. Stargan v2: Diverse image synthesis for multiple domains [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 8188-97.

[23] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]. International conference on machine learning, 2021: 8748-63. PMLR.

[24] Xian Y, Schiele B, Akata Z. Zero-shot learning-the good, the bad and the ugly [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4582-91.

[25] Xian Y, Akata Z, Sharma G, et al. Latent embeddings for zero-shot classification [C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 69-77.

[26] Li B, Qi X, Lukasiewicz T, et al. Controllable text-to-image generation [J]. Advances in neural information processing systems, 2019, 32.

[27] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [C]. International conference on machine learning, 2020: 1597-607. PMLR.

[28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:14091556, 2014.

[29] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019: 4401-10.

[30] Kwon G, Ye J C. Clipstyler: Image style transfer with a single text condition [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18062-71.

[31] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics [C]. International conference on machine learning, 2015: 2256-65. PMLR.

[32] Everaert M N, Bocchio M, Arpa S, et al. Diffusion in style [C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 2251-61.

[33] Wen L, Gao C, Zou C. CAP-VSTNet: content affinity preserved versatile style transfer [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18300-9.

[34] Qi T, Fang S, Wu Y, et al. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations [J]. arXiv preprint arXiv:240306951, 2024.

[35] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models [J]. arXiv preprint arXiv:210609685, 2021.

[36] 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): 600-12.

[37] Yoo J-C, Ahn C. Image matching using peak signal-to-noise ratio-based occlusion detection [J]. IET image processing, 2012, 6(5): 483-95.

[38] Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context [C]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014,

Proceedings, Part V 13, 2014: 740-55. Springer.
[39] Alexander K, Cyganiak R, Hausenblas M, et al. Describing Linked Datasets [C]. LDOW, 2009.
[40] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.