文章编号:1671-4598(2024)10-0243-07 DOI:10.16526/j.cnki.11-4762/tp.2024.10.035 中图分类号:TP399 文献标识码:A

基于扩散模型的印花图案生成方法设计

张佳伟[1], 李华军[1], 王秀丽[1,2], 朱 威[1,2]
(1. 浙江工业大学 信息工程学院, 杭州 310023; 2. 浙江省嵌入式系统联合重点实验室, 杭州 310023)

摘要: 印花图案的设计是服装生产领域非常关键的一环, 但目前人工设计的印花图案存在内容相似化、设计效率低下的问题, 因此设计并实现了一种基于扩散模型的印花图案生成方法; 采用深度学习技术提取、扩充现有印花图案数据集, 并从颜色和类别维度生成印花图案的文本描述, 完成印花图案数据集制作; 使用已制作的数据集微调扩散模型, 并对图像特征空间进行平铺处理, 使得生成的印花图案在边界过渡处满足纺织行业四方连续的要求。对局部扩散特性进行了分析, 结合文本引导, 实现细节可变的图文生图效果; 实验结果表明, 所设计的印花图案生成方法具备生成高质量印花图案的能力, 并且其特征空间平铺方法使得印花图案边界过渡处较为自然。

关键词: 印花图案; 深度学习; 多模态融合; 扩散模型; 图像生成

Design of Printed Pattern Generation Method Based on Diffusion Models

ZHANG Jiawei[1], LI Huajun[1], WANG Xiuli[1,2], ZHU Wei[1,2]
(1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China;
2. Joint Key Laboratory of Embedded Systems of Zhejiang Province, Hangzhou 310023, China)

Abstract: It is a crucial aspect to design printed patterns in the field of clothing production. However, manually designed patterns often have the shortages of content similarity and low design efficiency. Therefore, a printed pattern generation method based on a diffusion model was designed and implemented. This method employs a deep learning technology to extract and expand existing printed pattern datasets, generate the textual descriptions of these patterns from the dimensions of color and category, and complete the production of the dataset for printed patterns. The pre-made dataset is used to make small adjustments to the diffusion model and flatten the processing of the image feature space, which makes the generation patterns at the boundary transition edges meet the requirements of the textile industry. This paper analyzes the characteristics of local diffusion, and achieves a image generation effect with variable detail from images and text. Experimental results demonstrate that the designed pattern generation method is capable of producing high-quality patterns, and its feature space tiling method makes the transition of the printed pattern boundary more smooth.

Keywords: printed pattern; deep learning; multi modal fusion; diffusion models; image generation

0 引言

印花图案的设计是服装生产领域关键的一环, 但目前人工设计的印花图案存在内容相似化、设计效率低的问题, 如何快速设计制作新颖的印花图案, 成为印花图案设计行业的难题。

图像分割任务旨在将输入图像分割成多个具有语义信息的区域。UNet[1]是一种U型网络结构, 由一层一层堆叠的编解码器组成, 在能够保留高维度特征的同时, 恢复原始图像的分辨率, 进行有效的语义分割。显著性检测是识别图像中最引人注目的部分, 通常与图像中的对象、边界或者其他纹理结构相关, 可以看作是对图像分割任务的升级。TINet[2]采用了对称架构, 通过FGD编码器交互式地细化多级纹理和分割特征, 并加入边缘权重使模型更加关注边界处不易识别的像素过渡[3]。参考UNet的思想, 以每一个UNet网络为单位, 替换每层的编解码器, 提升了整个模型架构的深度, 从而达到了更好的显著性区域分割效果。

超分辨率任务通过增加图像的空间分辨率来提高图像的细节和清晰度。SRGAN[4]通过使用GAN[5]将超分得到的图像拟合到真实数据分布上, 避免了MSE损失导致的图像模糊性。ESRGAN[6]在SRGAN的基础上, 新增DDRB结构代替原有的残差结构, 并使用VGG[7]激活前的特征来计算感知域损失函数, 使超分结果边缘更加锐利, 符合视觉感知。Real-ESRGAN[8]在基本保持ESRGAN网络结构的情况下, 通过使用多阶的退化模型尽可能使合成数据贴近真实世界数据, 得到了非常好的超分辨率效果。

深度生成模型根据处理似然函数的不同方法可以分为3类: 近似法、隐式法和变形法[9]。近似法是通过变分或者抽样的方法求似然函数的近似分布, 主要包括变分自编码器VAE (Variational Auto-Encoder)[10], 其通过将噪声引入

收稿日期:2024-05-06; **修回日期**:2024-05-15。
基金项目:国家自然科学基金青年项目(62303414);浙江省自然科学基金探索青年项目(LQ23F030016)。
作者简介:张佳伟(1999-),男,硕士研究生。
通讯作者:朱 威(1982-),男,博士,副教授。
引用格式:张佳伟,李华军,王秀丽,等. 基于扩散模型的印花图案生成方法设计[J]. 计算机测量与控制,2024,32(10):243-249.

自编码器训练过程中，使图像的编码空间由从编码点变为一条连续的编码分布曲线，实现图像生成任务。VQ-VAE[11]在解码器之前加入了一个嵌入层，将离散的特征向量映射到一组特别的向量，提升了图像生成的质量，另外VQ-VAE在扩散模型领域作为特征编解码器同样拥有非常出色的效果。隐式法主要为Goodfellow等人提出的生成对抗网络（GAN），网络分为两部分，生成器（Generator）负责学习图像数据的分布，鉴别器（Discriminator）负责鉴别图像是否为生成器生成，这样互相对抗的训练过程可以让模型生成出更加逼真的图像，GAN优点在于通过对抗学习的方式为高纬度概率密度分布下的采样和训练问题提供了有效的解决措施[12]。变形法主要代表为扩散模型，Weiss等人[13]于2015年提出可以基于马尔科夫链特性，通过给图片加噪声得到训练数据，将一个已知分布变成另一个目标分布，并使用神经网络去学习这一过程，实现图像生成任务。Ho等人[14]在前者的基础上进一步改进，提出了去噪扩散模型DDPM（Denoising Diffusion Probabilistic Models），主要分为两个过程：前向加噪过程在真实的图像数据上逐步加入高斯噪声，后向去噪过程通过神经网络的方法去拟合数据分布。

Stable Diffusion[15]通过在潜在特征空间进行加噪去噪过程来降低计算复杂度，采用基于KL正则化的VAE编解码器，其下采样倍数为8，并且增加了感知损失以及基于patch的对抗训练，这可以减轻编解码带来的局部模糊，提升图像的整体质量；使用基于Transformer[16]的CLIP[17]的text-encoder模块作为文本条件编码器，其中patch为14，特征维度为768；去噪过程采用结合时序输入和交叉注意力机制的UNet结构，网络深度为4层，在前三层和最底层的连接层引入CrossAttention操作；在采样策略上，使用DDIM[18]来加速采样，其不再限制扩散过程必须是一个马尔科夫链，使得可以采用更小的步数来加速迭代去噪过程。

对于多模态大模型，数据就是核心驱动力。CLIP是用于对齐自然语言和视觉图像的跨模态模型，使用了4亿个从互联网搜集的图像-文本对数据集进行训练，取得了当时最好的多模态特征表征能力，其结构包含一个基于Text Transformer模型的文本编码器（Text Encoder）和一个基于ViT（Vision Transformer）[19]模型的图像编码器（Image Encoder），其将对应数据对作为正样本，错位数据对作为副样本，最大化正样本的余弦相似度，最小化负样本的余弦相似度，来优化文本和图像编码器，达到zero-shot图像分类的效果，并且因为其数据量和大模型参数量的优势，具有很好的泛化性，可以为很多下游任务提供文本编码或图像编码；而ChatGPT用到了惊人的3000亿单词语料数据进行训练。高质量、多样性和充分量级的数据集对于大模型的整体性能、泛化能力、鲁棒性等能力有很大的提升，但是目前印花图案领域的高质量图像数据集少之又少，光凭现有的数据集无法满足生成大模型的训练需求。

此外，由于纺织行业需要满足四方连续性质的印花图案来作为布匹生产的元数据，而对于通过文本描述控制生成结果的Stable Diffusion，即使加上“四方连续”、“可平铺”等文字提示，图像边界处仍然存在不自然的过渡，无法生成严格四方连续的印花图案。

因此，本文基于扩散模型设计并实现了一种印花图案生成方法。具体的，通过构建多模态的数据集自动化制作流程，进行印花图案数据集的制作，从前后景颜色和类别两个维度对印花图案进行文本描述。同时，利用已制作的印花图案数据集对预训练的Stable Diffusion模型进行微调，使其获得生成高质量印花图案的能力；通过在图像特征空间上进行平铺处理，使得印花图案在边界过渡处满足纺织行业四方连续的要求；主动对印花图案进行加噪，然后将带噪声图像作为扩散模型的初始图像进行迭代去噪，并通过文本引导，实现细节可变的图文生图效果。

1 印花图案数据集构建

1.1 现有数据集说明

现有印花数据为47 925张印花图案，其中23 215张印花图案带有类别描述，类别共9类：组合类花卉、几何图案、抽象图案、小碎花图案、纹理图案、中国风图案、大花朵图案、叶子图案及动物图案。除类别描述之外无其他文本描述，印花图案分辨率为512×512。由于网络传输、保存不当等原因，印花数据集中大部分保存格式为标准的PSD格式，其余为PNG格式。

现有模特衣物图共14 246张，基本从电商平台上获取，而在网络中流动的图像会遭受JPEG压缩、模糊、噪点等各种使图像受损的情况，因此整体清晰度较低，服装上的印花图案较为密集，平均分辨率为1 340×1 785。

1.2 模特衣物印花图案提取

由于现有的印花图案数据集较少，因此考虑从模特衣物图中提取印花图案纹理从而对现有数据集进行扩充，提取流程主要包含模特衣物分割和印花图案超分辨率两部分。

模特衣物分割采用SOLO[20]实例分割网络进行，避免一个模特图中出现多个衣物，训练数据集使用DeepFashion v2[21]，同时为了与模特图中的衣物相对应，将DeepFashion v2由原来的13类别改写为3类：上装（top），下装（bottom），裙子（dress），共计172 947张训练图片。并且观察到DeepFashion v2数据集存在衣物遮挡的问题，选取800张模特衣物图进行手动标注一同作为训练集。因此SOLO模型的训练过程如下：先通过DeepFashion v2数据集进行10轮训练，然后通过手动标注的模特衣物图进行二次微调，微调训练进行10轮，使得分割网络基本满足其应用场景。

使用DeepFashion v2单独训练和先使用DeepFashion v2训练再二次微调的分割网络做对比实验，得到的分割效果如图1所示，可以看到使用手动标注的模特衣物图进行微调训练后，衣物分割结果（图中黑色区域）基本不会被手、头发等遮挡物影响，使得后续印花图案提取更加准确无误。

由于模特图中印花纹理整体清晰度较低，无法直接使用分割结果作为印花图案数据集使用，因此需要对提取得到的纹理进行后处理以达到使用需求，使用Real-ESR-

图 1　衣物分割二次微调效果

GAN[14]超分辨率网络，超分倍数为 4，首先通过对现有数据集中的印花图案应用二阶退化模型，以模糊、下采样、加噪等图像处理，生成训练所需的数据集，然后在官方预训练模型的基础上进行微调，微调步数为 80 000 步，使超分辨率网络满足其使用场景。

模特衣物印花图案提取流程如图 2 所示：首先使用训练好的分割网络对模特衣物图进行分割得到衣物；然后计算衣服 mask 的最大内接矩形，使用大小为 128×128 和 512×512 的采样窗口在最大内接矩形中随机采样两次，以提取到不同尺度的印花图案，并且对捕捉到的纹理进行边缘阈值检测以过滤掉纯色图像；最后对 128×128 的低分辨率印花图案进行超分操作以提高其清晰度。

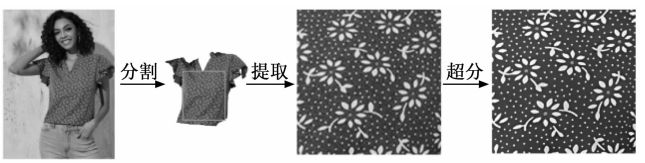

图 2　模特衣物印花图案提取流程

最终从模特衣物图中提取共 26 936 张印花图案以扩充现有印花图案数据集，最终扩充后的数据集共 74 861 张。

1.3　印花图案文本描述生成

从模特衣服上提取的印花图案不包含任何的文本描述，并且现有的印花图案数据集仅仅包含少量的类别描述，因此本节提出了一种基于颜色提取和 CLIP 类别分配的印花图案文本描述生成方法。

1.3.1　印花图案颜色描述

因为需要对颜色通过文本进行描述，所以根据比较常见的颜色词条对 HSV 色彩空间进行量化构建标准色板，具体操作如下：Hue 色相选取红、橙、黄、绿、青、蓝、紫等，然后将 Saturation 饱和度和 Value 明度进行配比，使得色相扩展成浅色系、原色、深色系等多类，例如浅红色、红色、深红色、朱砂红等，再加上白色到黑色渐变得到的灰色系，总计 136 种颜色，并对每种颜色进行命名，最终得到的标准色板中部分颜色如表 1 所示。

表 1　部分标准色板

Hex 值	中文名	英文名
＃FF1493	深粉色	DeepPink
＃87CEEB	天蓝色	SkyBlue
＃7FFFAA	碧绿色	Auqamarin
＃A9A9A9	深灰色	DarkGray

印花图案的颜色空间较为复杂，各形各异的颜色组合构成了印花图案，印花和背景颜色冗杂在一起，其前后景的边界较为模糊、不确定，如果直接对整个图像空间进行颜色描述并不能很好地去描述印花图案。同时观察到印花图案主要分为多层印花元素和一层背景元素，其中多层印花元素往往包含了不同的花朵、叶子等元素，可以视为同一层印花元素，因此使用显著性检测网 U^{2}Net 络将印花图案分为前景和后景两部分进行简化，分别进行颜色描述。利用现有印花图案数据集中部分分层的 PSD 格式文件构建前后背景数据集，将前后背景占比不均衡的印花图案去除，并且每个类别的图案均选取一部分，保证后续的解构效果，然后将最底层作为背景，其余层作为前景，生成掩膜图像，得到训练数据集，共计 22 527 张，训练步数为 400 000 步。

颜色描述生成步骤具体如下：通过显著性检测将印花图案分为前后背景两部分，并进行图像学腐蚀操作，从而减少边缘分割不准确和前后景交界处色彩渐变的影响。接着使用 K-Means＋＋聚类算法分别对前后景的 RGB 色彩空间进行颜色聚类，聚类中心数选取 5，得到总共 10 个图像的主要颜色。最后在 HSV 空间中计算聚类得到的主要颜色和标准色板的欧氏距离，采取百分比占比第一的和大于 25％的颜色作为印花图案的前后背景颜色描述。前后景分割效果和颜色描述生成结果示例如图 3 所示。

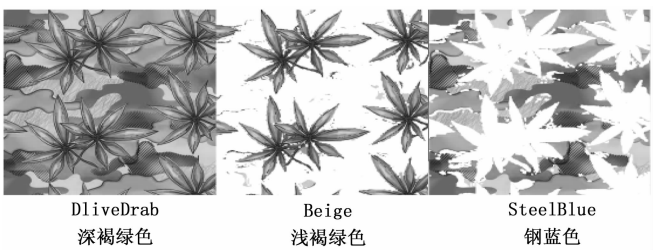

图 3　印花图案颜色描述

1.3.2　印花图案类别描述

因为印花图案数据集中有少量印花图案已经含有类别描述，所以同时考虑文本特征和图像特征，如图 4 所示，具体方法如下：使用 CLIP 文本编码器和图像编码器分别对印花类别库和已有类别的印花进行特征编码得到文本特征集合 $T_1 \cdots T_n$ 和图像特征集合 $I_1 \cdots I_n$，然后对待标注类别的图像进行编码得到图像特征 I，并计算图像特征之间的欧式距离以及文本特征与图像特征之间的余弦相似度，最终得到衡量印花图案与类别描述的距离，具体公式如式（1），

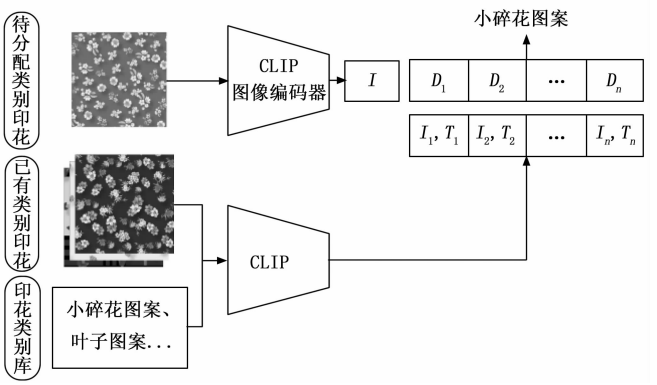

图 4 印花图案类别描述

然后根据阈值过滤并找到距离最近的类别，作为印花图案的类别描述。

$$D_i = \| I - I_i \|^2 + \frac{I \cdot T_i}{\| I \| \| T_i \|} \quad i \in [1,9] \qquad (1)$$

文本描述通过英文逗号“,”连接，并且因为多模态模型对于文本描述的权重与文本位置相关，头部权重较高，尾部较低，因此从重要程度考虑，将类别描述与提取得到的前景颜色进行组合放在首位，将背景颜色放在次位，例如：“深绿色的叶子图案，红色背景”。

2 基于扩散模型的印花图案生成

2.1 微调 Stale Diffusion

Stable Diffusion 网络结构如图 5 所示，虽然通过在潜在空间进行加噪去噪过程以节省计算资源，但因为其模型参数量较大、Transformer 结构计算密集型的特点，从随机初始化的参数开始训练至收敛依旧需要极大的计算资源和训练时间，因此一般对于 Stable Diffusion 模型采用微调训练的方式较为合理，使用特定领域的数据集使其从通用生成模型转变为专用生成模型，往往具有较好的效果。由于 VAE 编解码器和 CLIP 文本编码器皆通过大量的数据进行训练，具有较好的泛化表征能力，且图像的整体生成效果主要由去噪过程决定，因此此处冻结 VAE 和 CLIP 文本编码器，仅仅对去噪网络 UNet 进行微调训练。

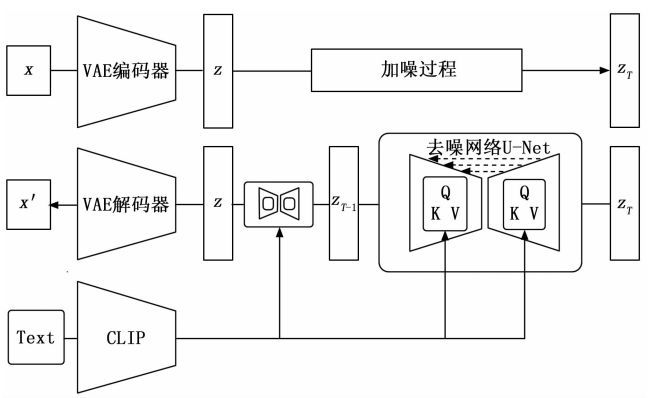

图 5 Stable Diffusion 网络结构图

微调训练的超参数如表 2 所示，预训练模型使用 stabilityai 的 stable-diffusion-2-base 版本。该模型首先在 LAION-5B 的 256×256 子集上从头开始训练 550k 步，该子集通过 punsafe 为 0.1 的 LAION-NSFW 分类器和审美分数大于等于 4.5 的过滤器进行筛选，然后又在分辨率大于等于 512×512 的同一数据集上进一步训练 850k 步。

表 2 Stable Diffusion 模型微调训练超参数

参数名	参数值
微调步数	150 k
微调分辨率	512×512
Batch size	1
初始学习率	1×10^{-5}

2.2 特征空间平铺

由于印花图案需要进行四方连续拼接来作为布匹生产的元数据，而 Stable Diffusion 模型生成的印花图案边界处存在不自然的过渡，像素值产生明显的突变，不满足四方连续的要求。

SeamlessGAN[22]将原始图像经过随机裁剪和编码器编码，然后在特征空间上做一次 2×2 平铺操作，通过一系列卷积和解码器，得到四方连续的纹理图案。由于 SeamlessGAN 是通过从原图中随机裁剪一个区域进行四方连续图生成，并且使用一个判别器去评估纹理过渡处的连续程度，如此不断地在原始图像上进行搜索，直至找到产生四方连续图像的区域，最终生成四方连续纹理图像，这本质上是轻量型网络在参数量不够多的情况下的一种妥协。

为了使生成的印花图案具有四方连续性质，参考并改进 SeamlessGAN[22]的做法，将 Stable Diffusion 模型中涉及图像生成部分的网络层全部进行特征空间平铺处理，具体操作如图 6 所示：在每一个卷积层之前，将特征图进行 3×3 平铺并裁剪中间部分，其中裁剪的大小由每个卷积层的 padding 决定，并且由于裁剪扩大了输入特征图，因此将对应卷积层的 padding 置为零，然后再通过卷积操作得到输出。由于特征空间平铺操作没有引入额外的参数，因此无需重新训练整个模型。

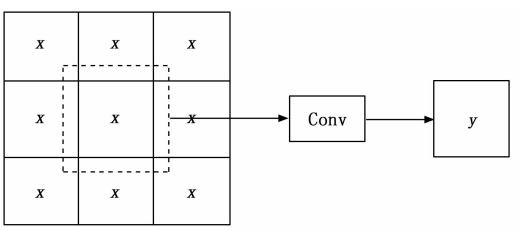

图 6 特征空间平铺区域

Stable Diffusion 模型即使加速生成过程，也需要多达几十步的采样步数，同时 UNet 编解码器有 4 层，并且前三层和连接层都会注入条件信息，所以扩散模型每一次迭代去噪、每一次经过 UNet 编解码器都会根据条件信息将特征空间平铺导致的突兀像素抹平，使其过渡处变得自然，即图 6 中虚线内部区域，这使得扩散模型可以直接生成四方连续的图像。

2.3　局部扩散特性

由于扩散模型的结构特点，图像生成的过程是从一个随机高斯噪声开始迭代去噪，直至得到清晰的图像，因此影响图像生成的变量只有通过交叉注意力机制注入的文本编码条件，所以生成图像仅包含了文字所描述的信息。

不同的真实图像之间虽然属于不同的分布，但是可以通过加噪过程让不同分布的支撑集越来越大，直到分布之间产生交集，这时从交集中的一点开始，使用训练好的扩散模型进行迭代去噪，就可以实现从一个分布变化到另一个分布，而原始的 Stable Diffusion 模型可以看作将图像完全噪声化，得到高斯噪声，与任意分布产生交集，再进行去噪生成。

就像盐结晶实验一样，给予一个小晶核，盐粒随着时间析出并生成一个大的盐结晶，如图 7 所示，可以先主动对印花图案进行加噪，然后将带噪声图像作为扩散模型的初始图像进行迭代去噪，同时输入文本描述信息引导模型生成印花图案，实现图文生图的任务。

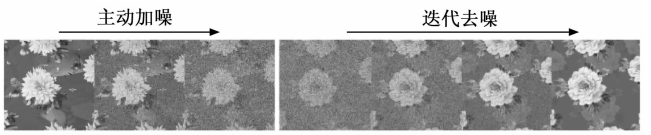

图 7　主动加噪再迭代去噪过程

3　实验设置与结果分析

3.1　实验环境

实验环境为 Linux 系统服务器，中央处理器使用第 12 代英特尔酷睿 i9-12900K，处理器线程数为 24，每个 CPU 有 16 个核心；图形处理器使用了 NVIDIA 的 GeForce RTX 3090 显卡，共有 10 496 个 CUDA 核心，显存大小为 24 G，NVIDIA 驱动版本为 525.147.05，CUDA 版本为 11.1；运行内存为 64 G。

3.2　印花图案文本描述生成实验

使用 1.3 节文本描述生成方法对印花图案进行文本描述，生成结果如图 8 所示，可以看到文本描述整体上符合印花图案的内容、色彩和风格。

最终得到的印花图案数据集共包含 74 861 个文本对，其类别分布如图 9 所示。

3.3　文生图实验

为了证明微调训练的有效性，进行 Stable Diffusion 模型文生图对比实验，实验结果如图 10 所示，其中左侧为原始模型的生成结果，右侧为微调后的生成结果，对应图像在生成时使用了相同的随机种子以避免随机性。可以看到，对于同一个文本描述，微调后的模型生成的印花图案更加符合文本描述，并且印花图案的整体质量和艺术性高于微调前。

3.4　特征空间平铺消融实验

为验证特征空间平铺操作的有效性，做消融实验：使用相同的文本描述和随机种子生成印花图案，并进行进行 2×2 平铺操作。

图 8　印花图案文本描述生成结果

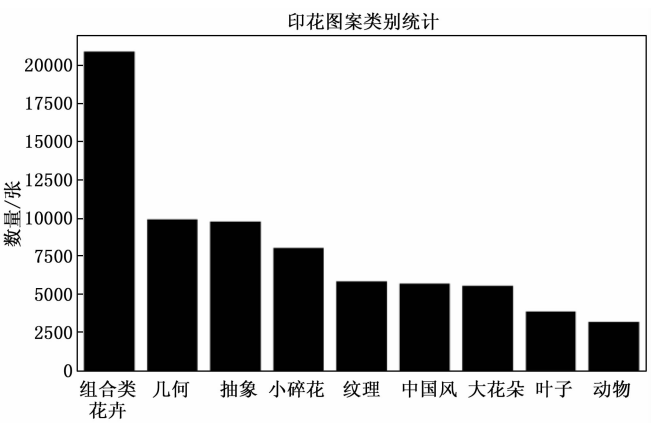

图 9　印花图案数据集类别分布

实验结果如图 11 所示，左侧为原始 Stable Diffusion 模型生成结果，右侧为加入特征空间平铺操作之后的模型生成结果，可以看到特征空间平铺操作使得印花图案边界处的像素过渡较为自然，满足四方连续性质。

3.5　图文生图实验

加噪过程首先破坏的是图像的高频特征，即细节纹理信息，然后才会破坏低频特征，即图像整体的语义信息。因此通常不会将语义信息全部破坏，否则扩散模型会从图文生图模型退化成文生图模型，同时如果加噪过少，则生成图像和输入图像一致，不会有什么变化。因此想要寻找

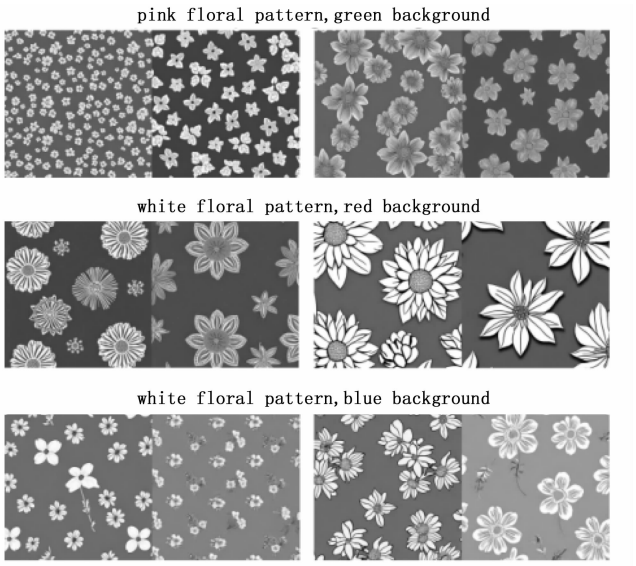

图 10 Stable Diffusion 模型微调前后对比

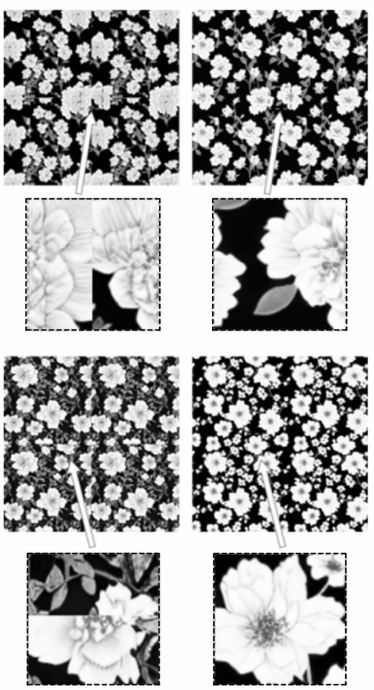

图 11 特征空间平铺消融实验结果

一个合适的加噪程度使得在保持印花图案整体语义和元素位置不变的情况下，改变图像的一些细节纹理。

对加噪程度为 40、60、80%的 3 种情况做横向对比实验以寻找合适的加噪程度，文本描述为“cartoon style”，即“卡通风格”。实验结果如图 12 所示，可以看到 40%加噪情况下生成图基本与原图一致；80%加噪情况下图文生图模型基本退化成为文生图模型，生成图与原图几乎没有关联；而 60%加噪情况下可以较好地保留原图的整体语义信息并且图像细节朝着文本描述的方向有一定的改变。

基于扩散模型的局部扩散特性进行图文生图实验，将原图叠加高斯噪声到 60%的程度，然后再进行迭代去噪过

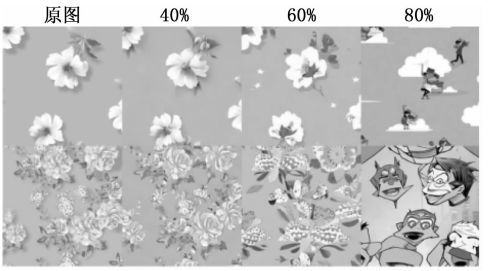

图 12 不同加噪程度下图文生图结果

程，同时通过文本描述引导图像生成，其中文本描述分别为“cartoon style”、“chinese style”、“style of painting”，即“卡通风格”、“中国风格”、“油画风格”，生成结果如图 13 所示，可以发现生成图像的风格能够根据提示词发生对应的改变，并且与风格迁移任务不同，生成图像的细节纹理也能朝着文本描述的语义发生一定程度的改变，实现细节可变的图文生图效果。

图 13 基于局部扩散特性的图文生图结果

4 结束语

本文设计并实现了一种基于扩散模型的印花图案生成方法。采用深度学习技术从模特衣物图中提取、扩充现有印花图案数据集，并从颜色和类别维度生成印花图案的文本描述。使用制作的数据集微调 Stable Diffusion 模型并对图像特征空间进行平铺处理，使其可以通过文本描述生成高质量、四方连续的印花图案，最后基于局部扩散特性，结合文本引导，实现细节可变的图文生图效果。

参考文献：

[1] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C] //Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, 2015: 234-241.

[2] ZHU J, ZHANG X, ZHANG S, et al. Inferring camouflaged

objects by texture- aware interactive guidance network [C] //AAAI Conference on Artificial Intelligenc-e (AAAI). 2021, 35 (4): 3599 - 3607.

[3] QIN X, ZHANG Z, HUANG C, et al. U2-Net: Going deeper with nested U-structure for salient object detection [J]. Pattern Recognition, 2020, 106: 107404.

[4] LEDIG C, THEIS L, HUSZáR F, et al. Photo-realistic single image super-resolution using a generative adversarial network [C] //International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 4681 - 4690.

[5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63 (11): 139 - 144.

[6] WANG X, YU K, WU S, et al. Esrgan: enhanced super-resolution generative adversarial networks [C] //European Conference Computer Vision (ECCV). Springer, 2018: 63 - 79.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J/OL]. (2014-09-04) [2024-03-15]. https: //arxiv. org/pdf/1409. 155-6v6. pdf

[8] WANG X, XIE L, DONG C, et al. Real-esrgan: training real-world blind super-resolution with pure synthetic data [C] //International Conference on Comp-uter Vision (ICCV). IEEE, 2021: 1905 - 1914.

[9] 胡铭菲，左　信，刘建伟. 深度生成模型综述 [J]. 自动化学报，2022，48 (1)：40 - 74.

[10] KINGMA D P, WELLING M. Auto- encoding variational bayes [J/OL]. (2013-12-20) [2024-03-15]. https: //arxiv. org/-abs/1312. 6114. pdf.

[11] VAN DEN OORD A, VINYALS O. Neural discrete representation learning [J]. Advances in Neural Information Processing Systems, 2017: 6306 - 6315.

[12] 张　彬，周粤川，张　敏，等. 生成对抗网络改进角度与应用研究综述 [J]. 计算机应用研究，2023，40 (3)：649 - 658.

[13] SOHL-DICKSTEIN J, WEISS E, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics [C] //International Conference on Machine Learning (ICML). ACM, 2015: 2256 - 2265.

[14] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [J]. Advances in Neural Information Processing Systems, 2020: 6840 - 6851.

[15] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C] //International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 10684 - 10695.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017: 5998 - 6008

[17] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C] //International Conference on Machine Learning (ICML). ACM, 2021: 8748 - 8763.

[18] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models [C] //International Conference on Machine Learning (ICML). ACM, 2021: 8162 - 8171.

[19] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [J/OL]. (2021-10-22) [2024-03-15]. https: //arxiv. org/pdf/2010. 11- 929. pdf.

[20] WANG X, KONG T, SHEN C, et al. Solo: Segmenting objects by locations [C] //European Conference Computer Vision (ECCV). Springer, 2020: 649 - 665.

[21] GE Y, ZHANG R, WANG X, et al. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images [C] //International Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 5337 - 5345.

[22] RODRIGUEZ-PARDO C, GARCES E. Seamlessgan: Self-supervised synthesis of tileable texture maps [J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29: 2914 - 2925.

(上接第 242 页)

[14] JOHNSON J, ALAHI A. Perceptual losses for real-time style transfer and super-resolution [C] //Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part II 14. Springer International Publishing, 2016: 694 - 711.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. Arxiv Preprint Arxiv: 1409. 1556, 2014.

[16] ZHANG Z, LI L, DING Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 3661 - 3670.

[17] WANG K, WU Q, SONG L, et al. Mead: A large-scale audio-visual dataset for emotional talking-face generation [C] //European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 700 - 717.

[18] KINGMA D P, BA J. Adam: a method for stochastic optimization [J]. Arxiv preprint arxiv: 1412. 6980, 2014.

[19] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13 (4): 600 - 612.

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

[21] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [J]. Advances in Neural Information Processing Systems, 2017: 30.