文章编号:1671-4598(2025)05-0313-09 DOI:10.16526/j.cnki.11-4762/tp.2025.05.037 中图分类号:TP399 文献标识码:A

基于多模态大模型的透明印花素材生成方法

李华军[1], 蒋俊豪[1], 金海云[2], 朱 威[1]
(1. 浙江工业大学 信息工程学院, 杭州 310023;
2. 杭州宏华数码科技股份有限公司, 杭州 310057)

摘要: 针对现有的图像生成模型无法满足纺织领域对生成多样透明印花素材专业要求的问题, 提出了一种基于多模态大模型的透明印花素材生成方法; 采用美学评分预测器构建高质量印花素材数据集, 并使用多模态大语言模型 BLIP3 进行数据集的标签语义生成; 通过多尺度分桶训练的方式微调 SD 模型, 并改进 VAE 模型将图像透明信息引入到图像生成空间中, 使得能够直接生成高质量的透明印花素材; 实验结果表明, 所设计方法在文生图、图生图两种模式下都能生成内容和风格多样的透明印花素材, 并且生成素材的边缘细节明显好于深度学习图像分割模型的结果。

关键词: 稳定扩散模型; 图像生成; 多模态大模型; 印花素材

Transparent Printing Material Generation Method Based on Large Multi-modal Model

LI Huajun[1], JIANG Junhao[1], JIN Haiyun[2], ZHU Wei[1]
(1. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)
2. Hangzhou Atexco Digital Technology Co., Ltd., Hangzhou 310057, China)

Abstract: Existing image generation models cannot meet the professional requirements of the textile industry for generating diverse transparent printing materials. To address this issue, a transparent printing material generation method based on large multi-modal model is proposed. Firstly, an aesthetic score predictor is used to construct a high-quality printing material data set. Secondly, the large multi-modal language model BLIP3 is employed to generate label semantics of the data set. Thirdly, the stable diffusion (SD) model is fine-tuned through multi-scale bucket training. Finally, the variational autoencoder (VAE) model is improved, and image transparency information is introduced into the image generation space, which can directly generate high quality transparent printing materials. Experimental results show that the designed method can generate transparent printing materials with diverse contents and styles in both text-to-image and image-to-image modes, and the generated material edge details of the model are significantly better than those of the deep learning image segmentation model.

Keywords: stable diffusion model; image generation; large multi-modal model; printing materials

0 引言

印花图案是设计师使用专业绘图软件将各种印花素材经过组合拼接、图像技法处理得到, 因此印花素材是印花图案创作的重要基础。为了提高创作效率和质量, 设计师通常对现有印花素材进行二次创作, 提取其主要元素, 并进行修改[1], 从而生成差异化的、各种风格的印花素材。传统的印花素材来源主要包括设计师手绘素材或者网络图片素材, 但是它们都存在一些局限性。手绘素材需要耗费大量的时间精力, 而网络素材质量一般不及手绘素材, 且可能存在版权纠纷问题。随着纺织生产规模的不断扩大和生产速度的加快, 自动生成印花素材成为行业迫切需求。近年来, 基于深度学习的图像生成算法取得了巨大成功, 为生成高质量的素材图案提供了可能。目前常用的图像生成网络有变分自编码器

收稿日期:2024-12-17; **修回日期**:2025-01-16。
基金项目:国家自然科学基金联合基金重点项目(U24A20270);杭州市重大科技创新项目(2022AIZD0077)。
作者简介:李华军(1999-),男,硕士研究生。
通讯作者:朱 威(1982-),男,博士,副教授,硕士生导师。
引用格式:李华军,蒋俊豪,金海云,等.基于多模态大模型的透明印花素材生成方法[J].计算机测量与控制,2025,33(5):313-321.

(VAE, variational autoencoder)[2], 生成对抗网络(GAN, generative adversarial network)[3], 以及最新的人工智能内容生成模型(AIGC, artificial intelligence generated content)。文献[4]提出了CycleGAN, 使得GAN能够完成图像到图像的生成问题, 且不需要配对的训练样本。LinkGAN[5]提出了一种易于使用的GAN训练正则化器, 它有助于将潜在空间的某些区域明确地链接到合成图像中的一组像素, 更方便地对GAN生成进行局部控制。尽管GAN网络在图像生成方面有着不错的效果, 但是在生成图像质量以及生成图像多样性等方面都不及Diffusion模型。基于Diffusion模型, 文献[6]提出了去噪概率扩散模型(DDPM, denoising diffusion probabilistic models), 通过从先验分布中抽样一个随机向量, 然后通过反向马尔可夫链进行祖先抽样, 从而生成新的数据点获得高质量的图像合成结果。文献[7]针对DDPM进行修改提出了IDDPM, 将DDPM中用常数指代的方差用模型学习, 将添加噪声的Schedule由线性改为余弦, 使得模型的样本质量和可能性可以随着模型容量和训练计算而平滑扩展。目前多模态大模型(LLM, large multi-modal model)在图像生成质量和灵活性方面获得了突破, 其中文献[8]提出了DALL.E模型, 将文本和图像Token作为单个数据流进行自回归建模, 该模型具有12亿个参数, 有着不错的生成效果。文献[9]提出了稳定扩散模型(SD, stable diffusion), 解决了计算资源有限问题, 在DDPM的基础上使用具有KL散度的VAE模型将图片从像素空间引入潜在空间, 去噪过程在潜在空间使用具有注意力机制的UNet结构进行图像生成, 保证速度的同时兼顾了图像生成质量。由于SD模型文生图模式无法准确控制生成图像的内容, 文献[10]提出了一种用于控制图像生成的神经网络架构ControlNet, 它冻结了大型扩散模型的主要结构, 并重用由数十亿图像训练主干中的深层和鲁棒编码层, 学习多样化的条件控制。

虽然目前已经有一些用于印花图案设计的生成模型[11], 其中文献[12]使用扩散模型生成整体的印花图案, 在视觉上取得了不错效果, 但存在着印花图案无法分层、分辨率有限等问题, 且生成图像不具备透明属性。当前主流的印花图案设计工作都是基于分层素材, 非常依赖具有透明度的分层元素来组成和创建内容。因此, 为了生成具有透明度的印花素材, 本文基于多模态图像生成大模型SD[9]和多模态大语言模型BLIP3[13]设计了一种透明印花素材生成方法, 能够通过输入文本或图像直接生成具有透明度的印花素材, 并且视觉质量满足印花图案设计师的要求。

1 DDPM模型与注意力机制

本文方法基于SD和BLIP3两种多模态大模型, 它们通过注意力机制来进行文本与图像之间不同模态的对齐, 保证图像特征能够与文本特征相匹配。其中最重要的多模态图像生成大模型SD生成图像是基于DDPM模型实现图像多样化生成, 解决生成图像单一化问题。

1.1 DDPM模型

去噪概率扩散模型DDPM[6], 通过对图像噪声分布的预测, 使模型能够生成形态各异的图片, 在图像生成领域取得了良好的效果。DDPM分为两个过程, 前向过程与后向过程。

前向过程是一个不断加噪声的过程, 式(1)是添加噪声的概率密度函数, 式(2)表示添加噪声后的图像:

$$q(X_t \mid X_{t-1}) = N(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I) \tag{1}$$

$$X_t = \sqrt{1-\beta_t}X_{t-1} + \sqrt{\beta_t}Z_t, Z_t \sim N(0, I) \tag{2}$$

其中: X_t 表示第 t 步添加噪声后的图像序列, 是每一步加噪的权重参数, $Z_t \sim N(0, I)$ 添加的高斯噪声, 随着 t 的变化不断增大, 满足 $\beta_1 < \beta_t < \beta_T (1 < t < T)$。式中 X_t 与 X_{t-1} 有着密切联系, 解出 X_t 需要知道前 $t-1$ 个噪声图像, 为了实现一步求解 X_t, 将其简化为式(3):

$$X_t = \sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}Z, Z \sim N(0, I) \tag{3}$$

其中: $\alpha_t = 1-\beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$。

反向过程与前向过程的区别在于, 反向过程需要利用前向扩散得到的高斯噪声进行训练, 预测噪声分布情况, 进而重建出真实图像。反向过程通过参数化的神经网络来估计每一步的去噪分布。反向过程的目标是从 X_T 得到 X_0, 已知 X_t 到 X_{t-1} 是添加随机噪声得到的, 因此利用贝叶斯公式, 在给定 X_t 的条件下, 可以计算前一时刻 X_{t-1} 概率, 也即从初始状态 X_0 推导到当前时刻的概率, 如式(4)所示:

$$P(X_{t-1} \mid X_t, X_0) = \frac{P(X_t \mid X_{t-1}, X_0)P(X_{t-1}, X_0)}{P(X_t, X_0)} \tag{4}$$

其中在不同时刻得到的噪声概率分布情况如式(5)所示:

$$\begin{aligned} P(X_t \mid X_{t-1}) &\sim N(\sqrt{\alpha_t}X_{t-1}, 1-\alpha_t) \\ P(X_t \mid X_t) &\sim N(\sqrt{\bar{\alpha}_t}X_0, 1-\overline{\alpha_{t-1}}) \\ P(X_{t-1} \mid X_0) &\sim N(\sqrt{\overline{\alpha_{t-1}}}X_0, 1-\overline{\alpha_{t-1}}) \end{aligned} \tag{5}$$

最后通过正态分布的概率密度函数形式来表示, 并将密度函数带入贝叶斯公式可得在 X_t 条件下, X_{t-1} 的概率分布如下:

$$P(X_{t-1} \mid X_t, X_0) \sim N\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha_{t-1}})}{1-\alpha_t}X_t + \frac{\sqrt{\alpha_t}(1-\alpha_t)}{1-\bar{\alpha}_t}X_0, \left(\frac{\sqrt{1-\alpha_t}\sqrt{1-\overline{\alpha_{t-1}}}}{\sqrt{1-\bar{\alpha}_t}}\right)^2\right) \tag{6}$$

通过式（6）可以推出所有时刻的噪声分布。反向过程中使用神经网络训练的目的是为了让噪声预测器 UNet 反向预测的噪声 $\overline{Z}$ 逼近前向过程添加的噪声 Z，随机噪声样本 X_T 通过预测噪声 $\overline{Z}$ 推导出 X_{T-1} 的条件概率分布，即 $P(X_{T-1} \mid X_T, X_0)$，对此条件概率分布进行随机抽样可以重建 X_{T-1} 的图像，接着重复上述过程即可得到原始图像 X_0。

1.2　注意力机制

在图像生成领域，需要文本和图像引导图像生成的过程中，注意力机制是必不可少的。文献［14］首次在循环神经网络（RNN，recurrent neural network）中使用了自注意力机制进行图像分类，能够通过自适应地选择一系列区域或位置，并以高分辨率处理选定区域来从图像或视频中提取信息。而 Transformer[15] 模型直接让注意力机制走向了巅峰，目前主流的大模型基本上架构都是基于 Transformer 和注意力机制。在 SD 中的模型中，UNet 模块使用了自注意力机制的同时也采用了交叉注意力机制，自注意力用于图像特征内部的信息聚合，交叉注意力用于让生成图像对齐文本，通过多注意力结合，实现文本控制图像生成。文献［16］揭示了扩散模型的交叉注意层和自注意层的潜在原因，提出了两种新颖的损失，以根据给定的空间布局在采样期间重新聚焦注意力图。

2　本文透明印花素材生成方法

在现有的印花素材生成方法中，生成图像的前景和背景是混叠一起的，无法直接生成只具有前景的图案，即不具备透明属性。印花素材实例如图 1 所示，其中透明印花素材（a）指的是具有透明背景的印花素材，（b）为（a）在专业软件上的效果，可以清晰看到素材与透明背景，（c）是在（a）中加入了其他背景，无法直接应用于印花图案创作，并且生成多余背景的印花素材是当前多模态大模型普遍存在的问题。

(a)分层印花素材　(b)分层印花素材(专业软件)　(c)带背景印花素材

图 1　印花素材

针对上述分层问题，本文设计的基于多模态大模型的透明印花素材生成方法，其网络结构主要分为 3 个模块：数据处理模块、图像编解码模块和印花素材生成模块。数据处理模块将 CLIP[17] 和多层感知机（MLP，Multilayer Perceptron）[18] 进行融合，通过 CLIP 模型提取图像的特征向量输入，并结合 MLP 网络对印花素材进行筛选，合理清洗数据集。图像编解码模块，对 SD 模型的 VAE 模块进行了改进，使其能够对具有透明背景的印花素材进行编解码，在编解码阶段分别处理图像的前景与背景，能够保证 VAE 模块不影响 UNet 模块的生成效果。印花素材生成模块采用了 DDPM 与 UNet 模块，通过不断迭代去噪，逐步生成出高质量透明印花素材。图 2 为本文方法整体网络架构。

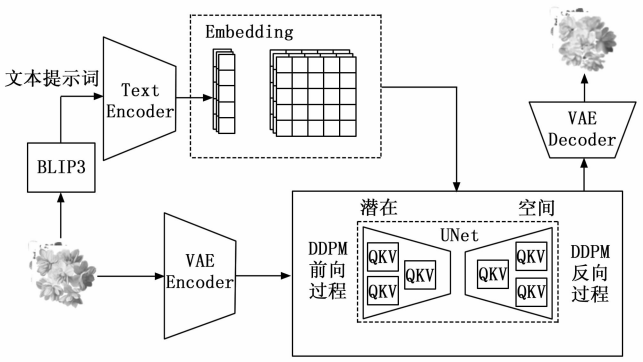

图 2　整体网络架构

2.1　数据处理模块

在深度学习图像生成领域中，数据集的质量是图像生成质量的基石。为了提高数据集质量，让数据分布均匀，用于后续模型的训练，需要对图像进行一系列前处理操作。数据集使用的是宏华数科授权使用的米绘数据集，有 24 万张印花素材图像，主要的风格标签共九类：现实风格、水彩风格、漫画风格、手绘风格、线条画风格、色块风格、中国画风格、彩铅风格、油画风格。除标签描述之外无其他文本内容描述，且全部图案均为具有透明度的 PNG 格式。但是由于图片质量参差不齐，存在质量不佳的图像会影响训练后的模型的生成效果，因此首先对图像进行质量筛选，选择高质量图像作为数据集。

本文方法采用融合 CLIP 与 MLP 模型的美学评分预测器，自动对数据集中的图片进行打分，其模型如图 3 所示。

其中 MLP 模型是 LAION 公司基于 AVA 数据集训练的，该数据集包含 250 000 张照片，每张图像都有 1.0 到 10.0 的美学评分，其中大多数美学评分为 2.0 至 7.0，将图像经过 CLIP 模型后得到的特征向量归一化后与美学评分一起送入 MLP 多层感知机中进行训练，其中在训练过程使用均方误差作为损失函数，通过评估真实评分和模型预测评分来不断调整 MLP 各层权重，从而使得模型预测评分不断逼近真实评分。预测器通过将高维特征向量输出为单个美学评分值。由于训练数据集的限制，该模型只能输出［1.0，10.0］区间内的评分值。本文方法中数据处理模块使用该美学评分预测器对印花素材数据集中的图像进行筛选，首先统计了数据集中所有图像的美学评分，其图像的美学评分在

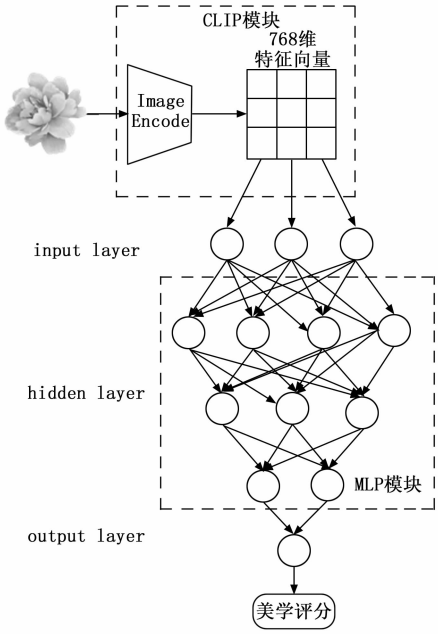

图 3 美学评分预测器结构

[2.0，7.0] 区间内，然后统计不同区间图像的分布范围，再通过多次比较不同范围内的图像，发现高于 5 分的图像具有优秀的主观视觉效果，最终从原始数据集中选取美学评分高于 5 分的图像，得到 15.8 万张图像作为训练数据集。

2.2 图像编解码模块

在图像生成任务中，图像的编解码质量往往会影响图片最后生成的质量，为了保证图像的生成速度与生成质量，本文方法采用了 VAE 模型作为编解码器，它能够对图像进行高效编码，将图像从像素空间编码到潜在空间，经过 DDPM 扩散过程后，能够从潜在空间重建一张和原图相近的图片。然而在 SD 模型中默认生成不具备透明背景的图像，单纯让 VAE 模型对透明印花素材编码，在网络中无法生成具有透明背景的印花素材，并且有可能导致生成的图像出现难以消除的伪影。为了让 VAE 模型重建出透明印花素材，本文方法在原始 VAE 模型上进行了改进，其改进 VAE 模型网络架构如图 4 所示。

引入额外的条件变量，即在潜在空间增加接收透明信息的 Alpha 通道。

Encoder 模块同时接收 RBG 通道和 Alpha 通道的信息，首先需要对具有透明信息的印花素材图像进行前处理，获取到图像的前景信息以及掩膜信息，再同时对这两个图像进行卷积操作，将其变为 128 维的特征向量，随后经过多个 DownEncoderBlock2D 块进行下采样，其中为了保证到达潜在空间之前保留较高的特征分辨率，可以使得潜在表示包含更多的细节信息，有助于解码器重建更精细的输出，最后一个 DownEncoderBlock2D 块

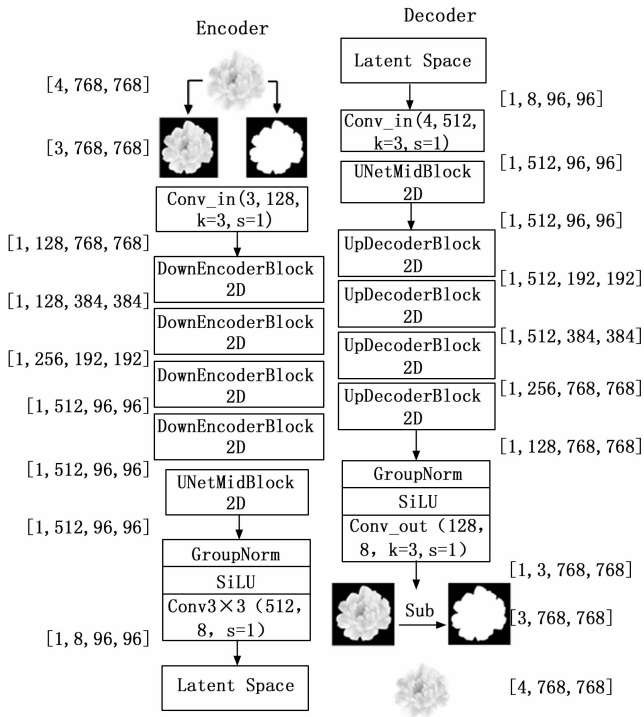

图 4 改进 VAE 模型网络架构

不进行下采样，再经过一个 UNetMidBlock2D 模块确保图像在潜在空间表示时能够保留尽可能多的细节和上下文信息，从而提高重建图像的质量，最后采用了一个组归一化 GroupNorm，非线性激活函数 SiLU 和 3x3 卷积结合，将特征向量编码到潜在空间，实现了高维图像空间映射降维到低维的潜在空间中，内存和运算量减小 64 倍。最终经过改进的 VAE 模型的图像能够正常重建出透明通道的印花素材。Decoder 模块是将在潜在空间中经过 DDPM 扩散后的特征向量进行解码，其中图像的掩膜信息，作为图像的偏移噪声用于后续透明图像生成，其中输入图像经过类似于编码过程，经过一系列上采样操作逐步还原成 RGB 图像，但是为了获取到具有透明通道的图像，还需要将 RGB 图像减去偏移噪声，最后就能得到带有透明通道的图像，同时还不会影响 UNet 模型内的噪声分布。

2.3 印花素材生成模块

为了使生成的印花素材多样化，并且能够生成出细节丰富的多尺度透明印花素材，本文方法在微调 UNet 模型的同时，采用了视觉多模态大模型 BLIP3 生成微调 SD 所需的图文对描述，并采用多尺度分桶微调的方法进行训练。

2.3.1 UNet 模块

目前微调 SD 有多种方法如文本反演[19]，DreamBooth[20]，低秩自适应方法（Lora）[21]等，这些微调方法都是用少样本来训练特定物品和风格，对噪声预测器

UNet 的扰动较小，但是对复杂多变，风格不一的印花素材而言效果不佳，并且重新训练 SD 需要极大的计算资源和图片数据，因此采用微调模型的方法进行训练，使用专业领域的数据集微调可以让模型生成特定领域图片。微调训练主要是针对 UNet 模块，改变它对印花素材的噪声分布预测情况，从而可以生成高质量的印花素材图案，其 UNet 模块如图 5 所示。

UNet 模块在潜在空间对 VAE 编码后的图片进一步编码，并进行 DDPM 前向加噪过程，在潜在空间对高斯噪声矩阵进行迭代去噪，每次去噪预测都受文本和时间步的引导，通过从随机高斯噪声矩阵中逐步去除预测的噪声，最终将该噪声矩阵转换为图像的潜在特征表示，其中文本即训练时需要的图片描述信息，推理时生成图片的内容描述信息，时间步为去除噪声的迭代步数。UNet 模块首先会对图像数据进行下采样，然后对数据进行上采样，在采样期间，为了防止信息丢失，下采样和对应的上采样之间会使用残差卷积连接。每个块之间都包含了两个部分，分别是残差卷积块和融合不同模态信息的 Transformer 块，同时时间步和其他约束等额外信息会输入到这两个模块中。在每一轮训练过程中，对于每个训练样本，关联一个随机选取的时间步长向量 t，该向量会被编码转化成对应的时间步长 Embedding 向量 E_t，将时间步长 t 对应的高斯噪声 Z 应用于原始图像，从而生成噪声图像。将时间步长 Embedding 向量 E_t 和噪声图像一起送入 UNet 网络训练，UNet 输出预测噪声 $\overline{Z}$，与实际高斯噪声 Z 进行比较，构建损失函数量化预测噪声与真实噪声之间的差异，逐步引导网络参数优化，损失函数如式（7）：

$$L = E_{x_0,Z,t}\left[\| Z-\overline{Z}(\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}Z,t)\|^2\right] \quad (7)$$

其中：t 为 $[1,N]$ 的均匀采样，$\overline{\alpha}_t=\prod_{i=1}^{t}\alpha_i$。通过使用印花素材图案微调后，模型的整体能力被重新校准，以更好地适应新的数据分布。

2.3.2　BLIP3 模型

SD 模型生成特定领域高质量图像必须微调模型来实现，但是公开的图文对数据集几乎没有与印花素材相关的。因此，针对印花素材图文对难以收集的情况，本文采用了一种多模态大模型 BLIP3 生成图文对数据集，BLIP3 模型如图 6 所示。该模型能够同时输入图像与文字，解决了多模态图文交错输入的问题，通过输入问题使模型准确地描述图片内容，生成符合要求的文字。相比于通用的大语言模型，BLIP3 能够识别图片内容，生成具有简洁明了特点的文本，能够理解印花图案的风格与内容，增强了 SD 模型对于文本内容的理解。BLIP3 模型将图像编码成 Visual Tokens，文本编码成 Text Tokens，最后按照顺序拼接起来送入大语言模型中生成关于图片的文本描述。并且在生成文本 Token 的能力上，BLIP3 远远优于传统的 CLIP 模型，其能够生成的 Token 数量远超 CLIP 模型。CLIP 模型只能提供 15 至 30 个有效 Token，而 BLIP3 模型可提供至少 50 个 Token，因此在训练时 BLIP3 模型提供的文本 Token 能让模型在微调中理解更多的信息，从而让文本更好的引导图像生成，达到生成图像更接近文本提示的效果。在模型训练时需要将图像与文本描述一起送入网络中，学习文本描述与图像内容的潜在关联，进而训练后的模型

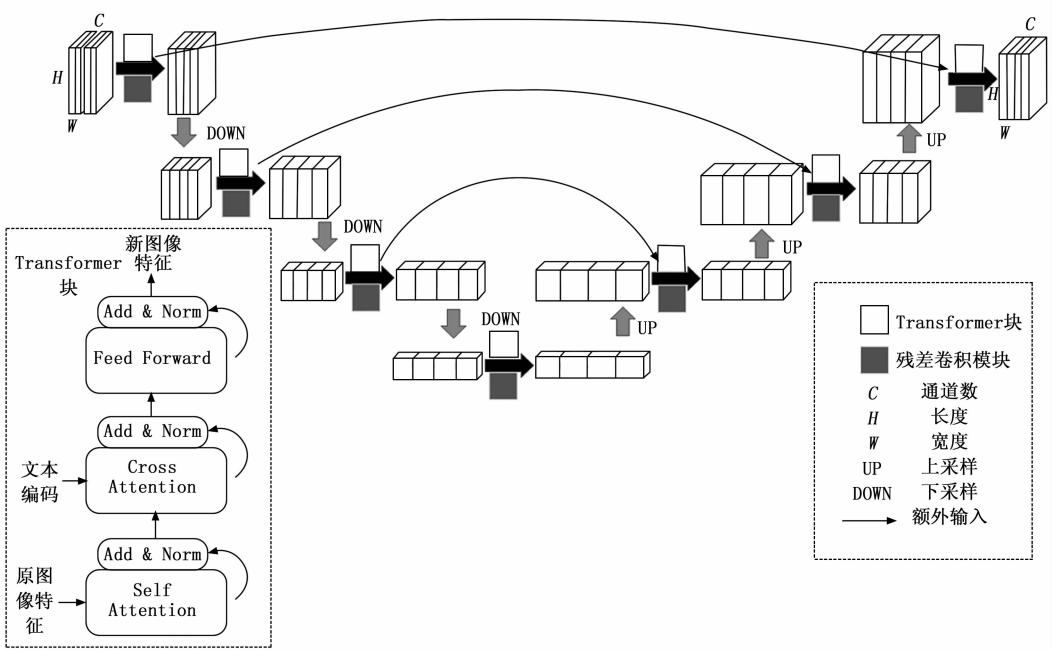

图 5　基于注意力机制的 UNet 模型

可以理解文本语义，按照文本描述生成内容。并且在图像生成过程，引入了 BLIP3 作为文本编码器，大大提升了模型对于文本的理解，使得图像在对于长文本描述时也能很好理解文字内容，生成符合描述的图像。

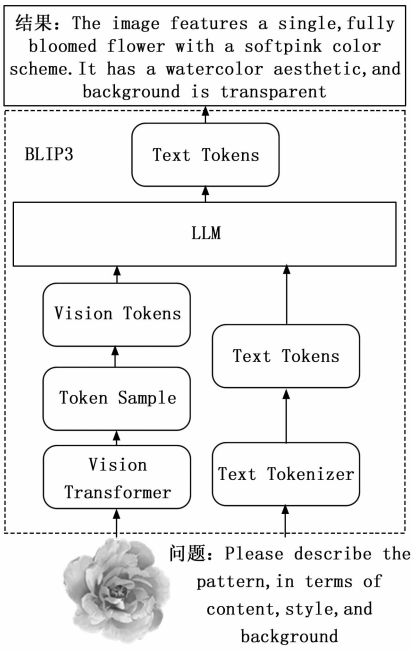

图 6 BLIP3 模型

2.3.3 分桶微调策略

预训练采用的 SD 模型 SDXL 版本，其模型训练时的基础分辨率为 1 024×1 024，在生成其他分辨率图像时会出现生成图像质量差，生成图像特征缺失的情况，并且由于数据集的分辨率大小不一，使用固定分辨率不能让模型学习到图像的长宽比分布信息，导致生成图像单一化。因此微调时采用了基于分桶的多尺度训练策略，即不同分辨率图片按长宽比进行分桶，训练时随机将桶中 BatchSize 大小数目的图片取出，可以增加模型的泛化能力，训练时数据集具体的分桶情况如表 1 所示。

表 1 分桶长宽参数

长	宽	长宽比
768	1536	0.5
768	1024	0.75
768	896	0.86
768	768	1.0
768	640	1.2
768	576	1.33
768	512	1.5

由于数据集大部分在 768 像素附近，选取 768 像素作为基值，缩放图片使其尽可能在 768 像素附近，同时相邻桶之间宽或高一般相差 64 像素左右。通过多尺度分桶微调可以保证印花素材的主体内容特征都能够被网络学习。

3 实验与结果分析

3.1 实验环境

在深度学习中由于 CPU 对训练和推理起到的作用十分有限，因此实验环境搭建主要依靠 GPU 服务器。由于微调多模态大模型需要显卡资源非常庞大，为了加快训练速度，训练时使用 80 G 显存的 NVIDIA H800 显卡，推理使用 24 G 显存的 NVIDIA RTX 4090 显卡。训练时采用的参数信息如表 2 所示。

表 2 微调模型参数

训练参数	参数信息
微调步数	100 k
微调分辨率	多尺度分辨率
Batch Size	4
梯度累计步数	4
学习率	0.000 001
学习率调度器	constant

3.2 文生图实验

文生图的原理是 SD 模型预先生成一个随机的高斯噪声图像，同时使用 CLIP 模型中 Text Encoder 部分将文本描述进行编码，生成特征矩阵，最后将特征矩阵、高斯噪声图像和生成步数等各种参数输入到 UNet 网络中对高斯噪声图像逐步去噪，最终通过 VAE 解码得到生成图像。为了保证实验的可靠性，将生成步数、随机种子、文本相关程度以及采样方式保持一致，输入不同文本描述验证模型的实际效果，文生图实验结果如图 7 所示。

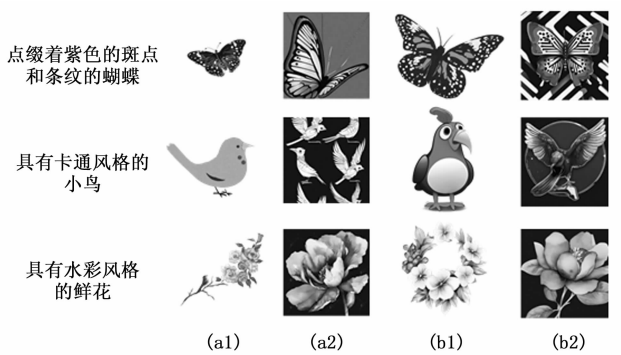

图 7 文生图实验结果

（a1）和（b1）是微调后的 SD 模型生成的透明印花素材，（a2）和（b2）是微调前模型生成的普通印花素材。从结果上来看，使用“点缀着紫色斑点与条纹的蝴蝶”作为文本描述，微调前的模型蝴蝶出现了部分主体内容缺失且背景杂乱的问题，而微调后的模型蝴蝶信息完整并且背景透明，并且另外两组微调后的效果同样

优于微调前的效果，因此微调后 SD 模型的生成图像整体质量高于微调前，并且包含了透明背景信息，可以更好地应用于印花图案创作

3.3　图文生图实验

图文生图的基本原理是在文生图的基础上，将模型随机生成的高斯噪声图像换成了实际输入图像，并对输入图像进行加噪处理，加噪强度决定了图像生成结果，加噪强度过大，图文生图会退化为文生图，加噪强度过小，会造成生成图像与实际输入图像基本一致。采用横向对比实验验证生成图像与噪声的关系，文本描述为“realistic style flower”，即真实风格的花朵，实验结果如图 8 所示。

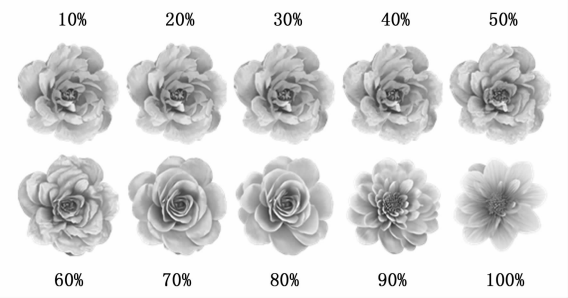

图 8　噪声强度对比实验

随着噪声强度增加，生成图像的信息发生了变化，在 50%以下噪声时，图像只有局部有着微小的变化，视觉基本无法分辨；在 50%～70%噪声时，图像的轮廓逐渐发生变化，且风格开始符合文本描述信息；70%噪声以上时，图像已经完全具备真实风格，且图像形态开始变化，与原始图像有较大变化，图像由原来的水彩画风格逐步变成文本描述的真实风格。相比于文生图，图文生图更具有可控性，在印花素材生成方面也更为简单，用户不需要去输入大量文本描述就可以生成与原图相似的图案。为了验证图文生图的效果，基于不同的文本描述进行图文生图实验，生成结果如图 9 所示。将噪声强度统一设置为 70%，文本描述设置为“watercolor style”“realistic style”“cartoon style”，即“水彩风格”“真实风格”“卡通风格”。通过图文生图实验发现，经过微调后的模型能够根据文本描述生成多种不同风格的图片，虽然图片风格发生了明显变化，细节纹理方面也出现了显著性变化，但是其内容主体依旧与原图保持一致，并且图案整体质量与原图一致，甚至部分效果优于原图。

3.4　客观指标

图像最终生成效果很大程度上取决于 VAE 模型的重建效果，通过对模型进行定量分析，本文方法将改进后的 VAE 模型与其他 VAE 模型进行了比较，评估其在图像重建与生成效果方面的性能。本文选取了 3 个常

图 9　图文生图实验结果

用的指标进行分析，分别是峰值信噪比（PSNR，peak signal-to-noise ratio）、结构相似性指数（SSIM，structural similarity）和感知图像块相似性（LPIPS，learned perceptual image patch similarity），实验结果如表 3 所示。

表 3　不同 VAE 模型重建比较

VAE 模型	PSNR	SSIM	LPIPS
VQVAE	27.542	0.768	0.181
SD2-VAE	30.145	0.766	0.133
SDXL-VAE	31.256	0.805	0.112
OURS	32.778	0.849	0.102

这些指标的计算均基于原始图像与 VAE 重建图之间的比较，以量化重建效果的差异，实验结果显示，在透明印花素材的重建结果中，改进后的 VAE 模型的 PSNR 为 32.778 dB，结构相似性指数 SSIM 为 0.849，感知图像块相似性 LPIPS 为 0.102，这些结果表明，改进后的 VAE 模型在透明印花素材的重建中，相较于原始 VAE 模型提高了图像重建的质量。

3.5　消融实验

为验证 VAE 模型改动以及微调训练的有效性，进行消融实验，使用相同的输入图像、文本描述与随机种子进行图文生图实验生成印花素材图案，实验结果如图 10 所示。

根据（a）（b）（c）（d）4 组图片的对比情况来看，在输入图像是透明印花素材时，改进前的 SD 模型生成的素材会出现边缘丢失的情况，以及生成异常素材的问题，且生成出来的素材质量不及原图。改进后的模型生成素材质量基本与原图相近，且不会出现边缘丢失的情况，同时能够直接生成透明印花素材，不需要后期进行二次处理，能够直接用作印花图案的创作。

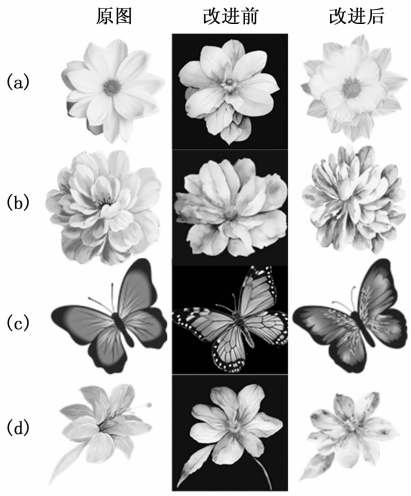

图 10 消融实验

3.6 对比试验

原始模型无法生成透明印花素材，但是可以通过使用分割模型对图像进行二次处理来生成透明图像。为验证微调模型与分割模型的效果，进行了对比实验。采用 3 个分割模型 U2-Net（U Square Net）[22]，SAM（Segment Anything Model）[23] 和 MatAny（Matte Anything）[24] 进行对比，实验将生成出来的透明印花素材加上黑色背景，让分割模型对图像进行分割操作取得前景后，再加入透明背景，即可得到透明印花素材，其实验结果如图 11 所示。

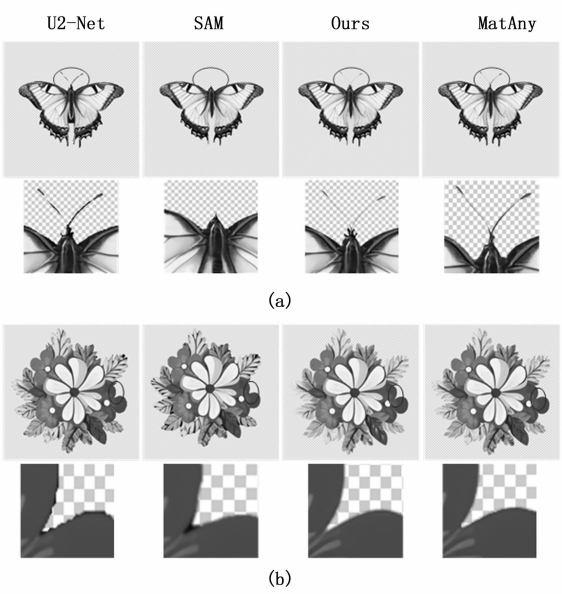

图 11 分割模型对比实验

在整体效果上分割后的图案与直接生成的透明图案差异不大，但是在边缘细节方面分割后的图像有明显问题，在（a）组图像中 U2-Net 模型分割后的图像在蝴蝶触角处会出现部分丢失问题，SAM 模型的蝴蝶触角则完全消失，MatAny 模型分割后在蝴蝶头部出现了丢失问题，（b）组图像中 U2-Net 模型花朵边缘出现大量不规则锯齿状像素，SAM 模型在两片花瓣中间存在大量黑色像素堆积，MatAny 模型边缘不够光滑且有少量锯齿状像素，整体来看，分割出来的图像在边缘与细节上的处理依旧存在问题。本文方法设计的透明印花素材生成算法能够直接生成高质量的透明印花素材，且能保证图像边缘完整不会缺失，有非常丰富的细节表现，因此证明了直接生成透明印花素材的效果优于二次分割后得到的透明印花素材。

4 结束语

本文方法设计了一种基于多模态大模型的透明印花素材生成方法。首先采用美学评分的方式筛选低质量印花素材。然后利用 BLIP3 多模态大语言模型制作数据集的图文描述。接着采用多尺度分桶的方式微调 SD 多模态图像生成模型，并在微调的同时改进 VAE 模型，同时结合 BLIP3 增强 SD 模型生成图像的文本特征，使其能够生成透明印花素材图案。实验从文生图、图文生图这两个角度验证了所设计方法的有效性，生成的素材能够达到设计印花图案的要求，并且在 PSNR、SSIM、LPIPS 等客观指标上有所提升。后续的工作将在解决局部细节问题、以及直接生成整体的分层印花图案等方面开展研究。

参考文献：

[1] 周 萌，葛小凡. 数码印花设计技术在纺织品图案设计中的应用研究 [J]. 艺术教育，2018，(21)：221－222.

[2] KINGMA D P，WELLING M. Auto- encoding variational bayes [EB/OL]. Arxiv Preprint Arxiv：1312. 6114，2013.

[3] GOODFELLOW I J，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems，2014，3：2672－2680.

[4] ZHU J Y，PARK T，ISOLA P，et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C] // International Conference on Computer Vision (ICCV)，2017：2223－2232.

[5] YANG Z，LI Y，ZHOU G. Ts-gan：Time-series gan for sensor-based health data augmentation [J]. ACM Trans on Computing for Healthcare，2023，4 (2)：1－21.

[6] HO J，JAIN A，ABBEEL P. Denoising diffusion probabilistic models [J]. Advances in neural information processing systems，2020，33：6840－6851.

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

[8] RAMESH A，PAVLOV M，Goh G，et al. Zero-shot text-to-image generation [C] // International Conference on Machine Learning (ICML)，2021：8821－8831.

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

[10] ZHANG L，RAO A，AGRAWALA M. Adding conditional control to text-to-image diffusion models [C] //International Conference on Computer Vision (CVPR). 2023：3836－3847.

[11] CAO H，TAN C，GAO Z，et al. A survey on generative diffusion models [J]. IEEE Trans on Knowledge and Data Engineering，2024.

[12] 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.

[13] XUE L，SHU M，AWADALLA A，et al. xGen-MM (BLIP-3)：A Family of Open Large Multimodal Models [EB/OL]. Arxiv Preprint Arxiv：2408. 08872，2024.

[14] MNIH V，HEESS N，GRAVES A. Recurrent models of visual attention [J]. Advances in neural information processing systems，2014，27.

[15] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale [EB/OL]. Arxiv Preprint Arxiv：2010. 11929，2020.

[16] PHUNG Q，GE S，HUANG J B. Grounded text-to-image synthesis with attention refocusing [C] //International Conference on Computer Vision and Pattern Recognition (CVPR)，2024：7932－7942.

[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). 2021：8748－8763.

[18] NIE Y，LI L，GAN Z，et al. Mlp architectures for vision-and-language modeling：An empirical study [EB/OL]. Arxiv Preprint Arxiv：2112. 04453，2021.

[19] GAL R，ALALUF Y，ATZMON Y，et al. An image is worth one word：Personalizing text-to-image generation using textual inversion [EB/OL]. Arxiv Preprint Arxiv：2208. 01618，2022.

[20] RUIZ N，LI Y，JAMPANI V，et al. Dreambooth：Fine tuning text-to-image diffusion models for subject-driven generation [C] //International Conference on Computer Cision and Pattern Recognition (CVPR)，2023：22500－22510.

[21] HU E J，SHEN Y，WALLIS P，et al. Lora：Low-rank adaptation of large language models [EB/OL]. Arxiv Preprint Arxiv：2106. 09685，2021.

[22] 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.

[23] KIRILLOV A，MINTUN E，RAVI N，et al. Segment anything [C] //International Conference on Computer Vision (ICCV)，2023：4015－402.

[24] YAO J，WANG X，YE L，et al. Matte anything：Interactive natural image matting with segment anything model [J]. Image and Vision Computing (IVC)，2024，147：105067.

(上接第 312 页)

[12] TANG Y，ZHAO Z J，LI C，et al. Open set recognition algorithm based on conditional gaussian encoder [J]. Mathematical Biosciences and Engineering，2021，18 (5)：6620－6637.

[13] 王小豪. 卫星通信系统中的干扰识别与决策技术研究 [D]. 西安：西安电子科技大学，2022.

[14] 李佳浩，杜子铭，周　博，等. 基于对抗互易点学习的无人机通信干扰开集识别方法 [J]. 信号处理，2024，40 (4)：639－649.

[15] HAN H，LI W，FENG Z B，et al. Proceed from known to unknown：Jamming pattern recognition under open-set setting [J]. IEEE wireless communications letters，2022，11 (4)：693－697.

[16] WANG G Q，GAO Y L，. Open-Set jamming pattern recognition via generated unknown jamming Data [J]. IEEE Signal Processing Letters，2024，31：1079－1083.

[17] TANG Y，ZHAO Z J，CHEN J，et al. Open world recognition of communication jamming signals [J]. China Communications，2023，20 (6)：199－214

[18] ZHOU Y，SHANG S，SONG X，et al，Intelligent radar jamming recognition in open set environment based on deep learning networks [J]. Remote Sensing，2022，14 (6220)：6220.

[19] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition [C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016：770－778.

[20] MILLER D，SUNDERHAUF N，MILFORD M，et al. Class anchor clustering：A loss for distance-based open set recognition [C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision，2021：3570－3578.

[21] 胡　鸿. 典型信号调制方式开集识别算法研究 [D]. 成都：电子科技大学，2024.