基于深度学习的虚拟人物形象生成和设计研究

梁涛[a]，姚怡然[a]，丁满[b]
（河北工业大学 a.人工智能与数据科学学院 b.建筑与艺术设计学院，天津 300132）

摘要：目的 为实现虚拟人物形象与用户情感偏好的匹配，提高虚拟人物形象设计效率，提出基于深度学习方法的虚拟人物图像生成方法。**方法** 首先收集相关风格图片并制作数据集，用于初始模型的微调；其次确定用户对虚拟形象的情感偏好与细节要求，得到生成所需的文本提示词；再次通过文本生成的方式迭代生成目标形象并初步筛选符合要求的形象；最后通过图像生成对形象进行局部优化和调整，使其符合用户情感目标。**结果** 以虚拟人物形象为研究对象，运用基于 Stable diffusion 模型的文本和图像两种生成方法得到符合用户情感需求与偏好的虚拟人物形象产品，并可快速调整人物形象的细节特征。**结论** 通过实例验证，所提方法可快速、高质量生成符合用户情感目标的虚拟人物形象，极大地提高虚拟人物形象设计的效率，丰富设计方法。

关键词：虚拟人物形象；形象生成；用户情感；深度学习；文本生成图像
中图分类号：TB472 **文献标识码：**A **文章编号：**1001-3563(2023)16-0059-08
DOI：10.19554/j.cnki.1001-3563.2023.16.007

Generation and Design of Virtual Characters Based on Deep Learning

LIANG Tao[a], *YAO Yi-ran*[a], *DING Man*[b]
(a.School of Artificial Intelligence, b.School of Architecture & Art Design,
Hebei University of Technology, Tianjin 300132, China)

ABSTRACT: The work aims to propose a method of virtual character generation based on deep learning method, in order to achieve the matching of virtual character and users' emotional preference and improve the efficiency of virtual character design. Firstly, relevant style images were collected and a dataset was created for fine-tuning the initial model. Then, the user's emotional preferences and detail requirements for the virtual character were determined to get the text cue words required for generation. Next, the target image was generated iteratively by text generation and the images in conformity with the requirements were screened initially. Finally, the character was locally optimized and adjusted through image generation to make it meet the users' emotional objectives. With the virtual character as the research object, both text and image generation methods based on the Stable diffusion model were used to obtain virtual character products conforming to users' emotional needs and preferences, and could quickly adjust the detailed features of the character. The proposed method is verified to be able to generate high-quality characters that meet users' emotional objectives quickly, which greatly improves the efficiency of character design and enriches the design methods.

KEY WORDS: virtual character; image generation; user emotion; deep learning; image generation from text

元宇宙概念的热门加速推动了虚拟人产业的发展，目前已在电商行业、直播领域、社会服务等领域取得了成功[1-4]。虚拟人物形象是虚拟人在视觉上的表现形式，能够反映虚拟人的个性、情感和社会属性[5,6]。用户往往希望通过选择或创造出与真实输入相似的虚拟人物形象，增强自身的归属感与情感表

收稿日期：2023-04-11
基金项目：国家自然科学基金（52275243）
作者简介：梁涛（1975—），男，博士，教授，主要研究方向为新能源大数据分析、深度学习与人工智能。
通信作者：丁满（1979—），女，博士，教师，主要研究方向为产品色彩情感设计、智能设计。

达[7-9]。然而，在虚拟人产业欣欣向荣的背后，虚拟人物形象的高制作门槛限制了行业的发展，因为无论是 AI 生成的虚拟人模型，还是由专业画师和建模师制作的虚拟人物模型，都需要消耗大量的财力物力[10-12]。而图像生成作为计算机视觉领域近年来发展最迅猛的领域之一，特别是性能出色的 Stable diffusion 模型，可以为虚拟人物形象设计提供更多的可能性与发展空间。因此，本文基于 Stable-Diffusion 模型的文本、图像生成图像功能，通过文本提示词约束、逐次迭代训练生成高质量、多样化的虚拟人物形象，以期满足用户个性化情感需求，提高虚拟人物形象设计的效率与质量。

1 深度学习概述

近年来由于计算机视觉与领域的迅速发展，关于图像生成的研究收到了许多学者的关注。王晓红等[13]基于生成对抗网络提出了一种提取书法图像特征并自动生成风格化书法图像的方法，用于修复书法图像中的残缺字体。张磬瀚等[14]为了解决文物图像的不易保存和物理方法修复困难等问题，提出一种基于生成对抗网络的图像修复算法进行文物图像残缺部分的生成。王晓慧等[15]将 AI 画作生成技术等图像生成技术应用于个性化文化创意产品的设计，为文化创意产品设计提供了新的思路。在图像生成领域中 Stable-Diffusion 模型通过在潜在空间完成迭代去噪操作，使图片生成稳定性和质量都大幅提高。因此，本文拟基于 Stable-Diffusion 模型研究虚拟人物形象设计方案，在满足用户个性化情感需求同时，提高虚拟人物形象设计的效率与质量。

1.1 Stable-Diffusion 模型

Stable-Diffusion 是基于潜在扩散模型（Latent Diffusion Models，LDMs）改进的文本生成图像（text-to-image）模型[16-18]。该模型通过在一个潜在表示空间中迭代含有噪声的数据生成图像，然后将表示结果解码为完整的图像。早期的扩散模型在 AI 绘画中效果不好，单张图生成需要 10~15 min，而 Stable- Diffusion 模型则通过在潜在空间而不是像素空间来执行迭代去噪操作，使图片生成稳定性和质量都大幅提高，生成速度提高了 100 倍，相比之前的扩散模型生成所需要 600~900 s 一张，只需要 6~10 s 即可生成一张图片。Stable-Diffusion 模型的提出，也促进了文本生成图像领域的热门，文本生成图像和图像生成图像见图 1。

相比于 LDMs 采用 laion-400M 数据训练，Stable-Diffusion 模型在规模更大的 laion-2B.en 数据集上进行训练，在使用更多的训练数据的同时，采用数据筛选来提升数据质量，如去掉有水印的图像及选择美学评分较高的图像。LDMs 采用随机初始化的 transformer 作为文本编码器，Stable-Diffusion 模型采用的是预训练好的 CLIP 文本编码器进校编码输入文本，而采用预训练的文本模型效果往往要优于从零开始训练的模型。LDMs 只在 256×256 分辨率上训练，而 Stable-Diffusion 则先在 256×256 分辨率上预训练，然后在 512×512 分辨率上进行微调，因此 Stable-Diffusion 在更大分辨率的图片生成任务中表现更加出色。

Stable-Diffusion 模型主要有三个组件：Clip Text 用于文本编码；U-Net 网络+调度函数，用于逐步处理或扩散信息到潜在表示空间；自编码中的解码器部分，用于绘出最后输出的图像。

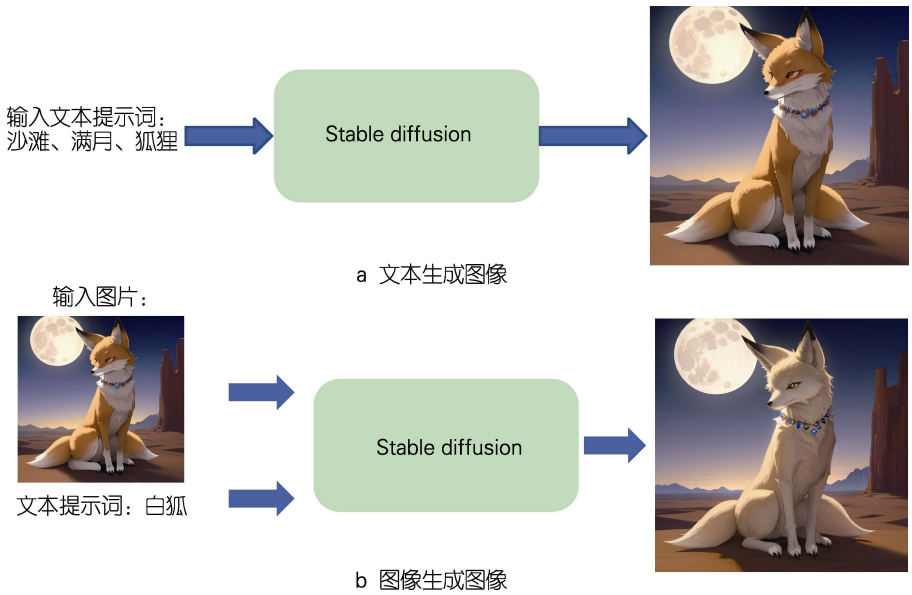

图 1 Stable diffusion 模型生成图像示意图

Fig.1 Schematic diagram of image generation by Stable diffusion model

1.2　CLIP 模型

Open AI 公司的 Alec Radford 等在 2021 年提出的 CLIP（Contrastive Language-Image Pre-training）模型[19-22]，是通过自然语言监督学习视觉特征的。CLIP 模型是先利用从互联网上获取的大规模图文匹配对进行训练，然后再应用到计算机视觉任务上。而常规的图像分类模型通常基于有类别标签的图像数据集进行全监督训练，例如在 Imagenet 上训练的 Resnet 网络，在 JFT 上训练的 ViT 等。这种训练方法对数据量的需求非常高，需要大量人工进行标注。同时模型的适用性和泛化能力受到限制，不适于迁移学习。而 CLIP 模型没有预先定义标签类别，直接利用从互联网获取的上亿数量的图像-文本对进行图文匹配任务的训练，将文本作为图像的标签，即利用自然语言作为监督信号来学习视觉特征，目前已经成功迁移应用于 30 个计算机视觉任务中。此外，CLIP 无需利用 ImageNet 的数据进行训练，即可达到和 ResNet-50 在该数据集上有监督训练的结果。

1.3　U−Net 模型

Ronneberger 等人[23-25]在 FCN 网络基础上，于 2015 年提出了 U-Net 网络，并在当年 ISBI 挑战赛的细胞分割任务中测试了 U-Net 网络的性能，证明 U-Net 网络在数据量较小的情况下，仍旧能获得精确的分割结果，并实现对图像的像素级分类。其凭借着出色的性能被广泛应用于语义分割的各个方向。

U-Net 网络由编码器、解码器和跳跃连接三个部分组成，因网络模型呈 U 形对称结构，故取名为 U-Net 网络。该模型延用 FCN 影像语义分割的思想，即先采用卷积层、池化层提取特征，然后采用反卷积层还原影像位置信息。U-Net 模型运算由编码部分与解码部分构成，压缩通道有以一系列编码器逐层提取影像的特征，随着层数加深，提取的特征具有更高层次的语义信息。扩展通道借助一系列解码器还原影像的位置信息，有助于模型整合底层、中层和高层特征[26-28]。

2　基于深度学习的虚拟人物形象生成和设计流程及方法

2.1　设计流程

本文基于深度学习进行虚拟人物形象生成与设计，按是否有初始输入的人像图片分为基于文本生成图像和基于图像生成图像两部分内容，具体流程见图 2。

2.1.1　基于图像的虚拟形象生成和局部设计方案

1）确定用户对虚拟形象的情感偏好与细节特征，如发色、背景、衣服、人物姿态，完成用户的情感需求分析，并得到后续生成图片所需的文本提示词。

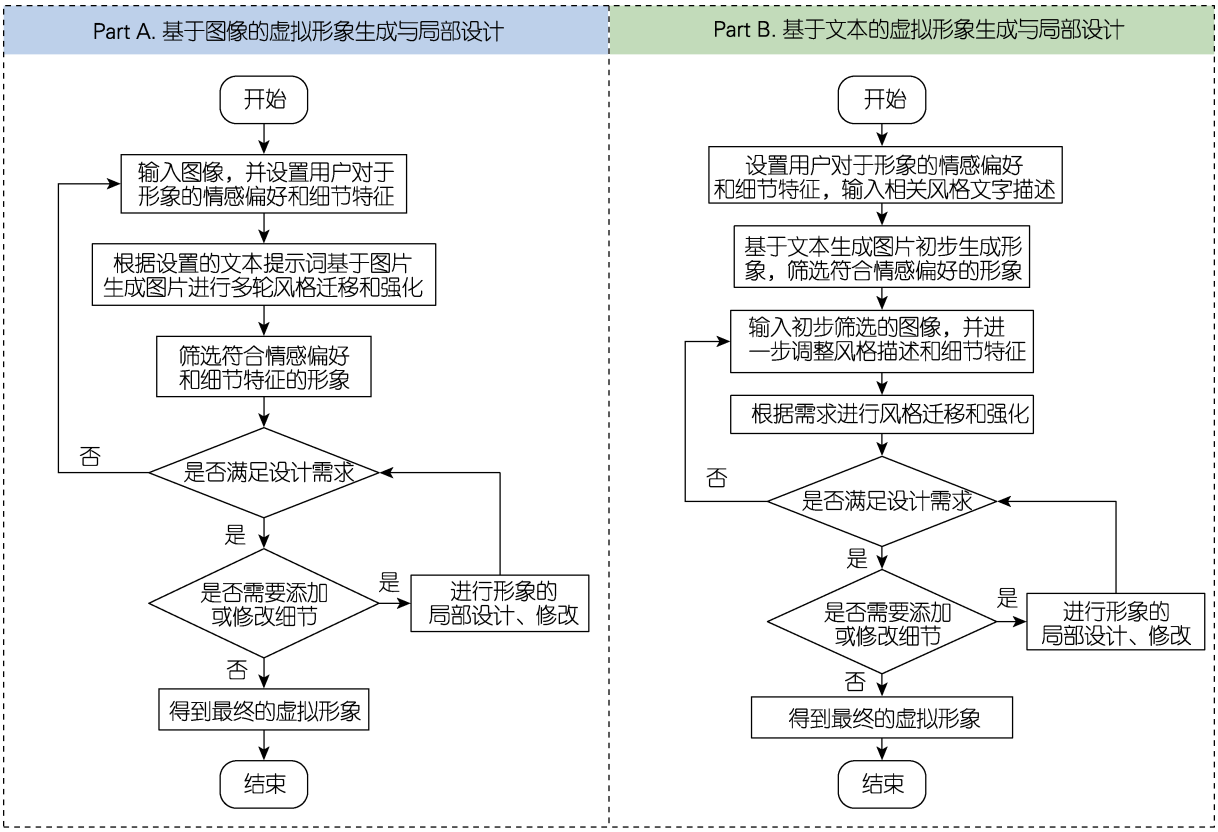

图 2　基于文本、图像的虚拟形象生成与局部设计

Fig.2 Virtual image generation and local design based on text and images

2）将图片输入至图像生成图像网络中，设定文本提示词和网络模型参数用于虚拟形象生成，逐步测试迭代生成目标风格的虚拟形象，最后筛选出与用户情感偏好相关度最高的虚拟形象。

3）得到相关度最高的虚拟形象后，如果想要添加局部效果使整体效果更加和谐美观，可将虚拟形象的图片和指定的局部区域，以及局部区域的文本提示词输入至图像局部修正的模型中，从逐步迭代的结果中挑选出符合要求的新虚拟人物形象。

2.1.2 基于文本的虚拟形象生成和局部设计方案

1）确定用户对虚拟形象的情感偏好与细节特征，如发色、背景、衣服、人物姿态，完成用户的情感需求分析，得到后续生成图片所需的文本提示词。

2）确定网络模型参数并输入正向和反向的文本提示词，通过文本生成图像网络得到一批图片，从输出图片中初步筛选与用户情感偏好相关度最高的虚拟形象图片。

3）将第“2）”段筛选的图片继续输入至图像生成图像网络中，继续设定文本提示词和网络模型参数用于图片细节和风格强化，逐步测试迭代生成目标风格的虚拟形象，最后筛选出最符合用户预期情感目标的虚拟形象。

4）得到最符合设定情感目标的虚拟形象后，如果想要添加局部效果，以使整体效果更加和谐美观，可将虚拟形象的图片和指定的局部区域，以及局部区域的文本提示词输入图像局部修正的模型中，从逐步迭代的结果中挑选出符合要求的虚拟人物形象。

2.2 相关算法设计

2.2.1 Stable-Diffusion 模型结构

Stable-Diffusion 模型整体框架如图 3 所示，从输入文本到输出的图像大致需要经过三个步骤。

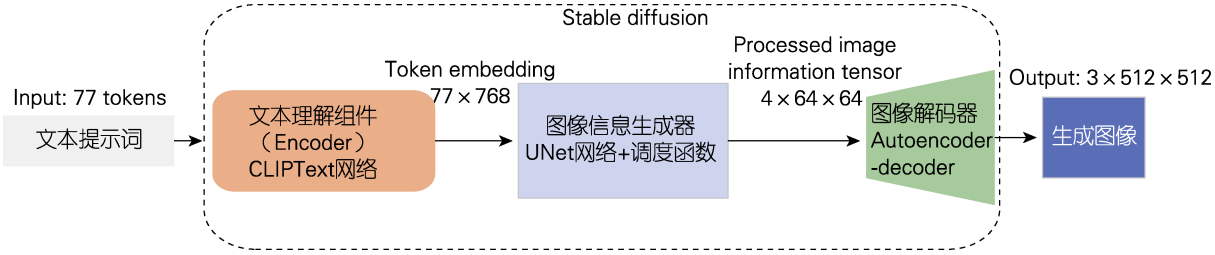

图 3 基于 Stable-Diffusion 的虚拟人物形象生成模型结构图

Fig.3 Structural diagram of virtual character generation model based on Stable-Diffusion

第一步，文本提示词编码过程。模型将潜在空间的随机种子和文本提示词同时作为输入，然后使用潜在空间的种子生成像素大小为 64×64 的随机潜在图像表示，通过 CLIP 的文本编码器将输入的文本提示词转换为大小为 77×768 的文本嵌入。文本中的每一个词都有会对应一个长度为 768 的向量特征。

第二步，使用 U-Net 进行扩散过程。使用含注意力机制的优化 U-Net 模型，在接受文本嵌入计算注意力的同时迭代地对随机潜在图像表示进行去噪操作。U-Net 的输出是噪声的残差，用于通过调度算法计算去噪的潜在图像表示。调度算法根据先前的噪声表示和预测的噪声残差计算预测的去噪图像表示。这样可以逐步检索更好的潜在图像表示。Stable-Diffusion 中涉及的调度算法包括 PNDM 调度、DDIM 调度+PLMS、K-LMS 调度等。

第三步，通过 VAE 对潜在图片进行解码。解码器根据图像信息生成器输出的信息绘制图像，其只需在结束时运行一次，即可生成最终像素图像，完成图片输出。

凭借扩散机制，Stable diffusion 也能完成各种图像到图像（图片修复，图片超分）和文本到图像的任务。

2.2.2 CLIP 模型结构

CLIP 的模型结构包括两个部分，即文本编码器（Text Encoder）和图像编码器（Image Encoder)。文本编码器选择的是 Text Transformer 模型；图像编码器可以在两种模型中选择，一是基于 CNN 的 ResNet，二是基于 Transformer 的 ViT。其训练过程见图 4。

训练过程分为三个步骤：

1）假设一个训练批次中包含了 N 个（文本-图像）对，将这 N 个文本先通过文本编码器进行编码。假设文本编码器将每条文本编码为一个长度为 d_t 的一维向量，那么这个训练批次的文本数据经过文本编

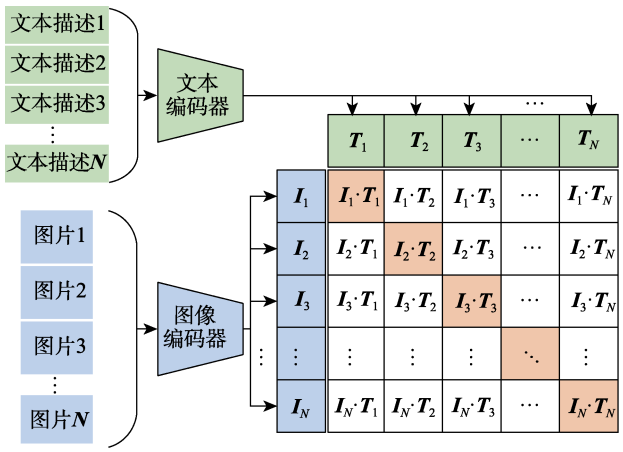

图 4 CLIP 训练过程

Fig.4 CLIP training process

码器输出为$[T_1,T_2,\cdots,T_N]$，维度为(N,d_i)；同理将这 N 个图像通过图像编码器进行编码，假设图像编码器将每条文本编码为一个长度为 d_i 的一维向量，那么这个训练批次的图像数据经过图像编码器输出为，$[I_1,I_2,\cdots,I_N]$维度为(N,d_i)。

2）得到的$[T_1,T_2,\cdots,T_N]$和$[I_1,I_2,\cdots,I_N]$中，文本-图像是一一对应的，例如 T_1 与 I_1 对应，T_2 与 I_2 对应，…，T_N 与 I_1 对应，这 N 个对应关系记为正样本；而不能对应的文本-图像标记为负样本，例如T_1与I_2不对应，T_N与I_{N-1}不对应。因此，一共有 N 个正样本，N^2-N 个负样本。这样正负样本就可以作为正负标签，用来训练文本编码器和图像编码器。

3）通过计算 I_i 与 T_j $(i,j\in[1,N])$之间的余弦相似度 $I_i\cdot T_i$，来度量相应的文本与图像之间的对应关系。余弦相似度越大，表明 I_i 与 T_j 的对应关系越强，反之越弱。那么训练任务便转换成通过训练文本编码器和图像编码器的参数，最大化 N 个正样本的余弦相似度，最小化 N^2-N 个负样本的余弦相似度。如图 5 所示，即最大化对角线中橙色块的数值，最小化其他非对角线的数值。优化目标见式（1）。

$$min\left(\sum_{i=1}^{N}\sum_{j=1}^{N}(I_i\cdot T_j)_{(i\neq j)}-\sum_{i=1}^{N}(I_i\cdot T_i)\right) \tag{1}$$

CLIP 模型在文本编码和图像编码部分采用现有编码模型的同时，引入对比学习思想，在新构建的大规模文本-图像数据集（WebImageText，WIT）上进行预训练，成功的将图片摘要预测任务转变为文本和图像匹配的预训练任务，并取得了优异的结果。CLIP 模型使用经典的双流编码结构在编码过程中没有进行不同模态之间的交互，而是使用编码得到的特征直接计算相似性以完成匹配预测。CLIP 有着良好的迁移能力，具有可以无需微调网络参数直接应用到下游任务的特点。

2.2.3　U-Net 模型结构

U-Net 网络模型结构如图 5 所示，由左侧压缩通道（编码部分）和右侧扩展通道（解码部分）构成。编码部分原理与卷积神经网络同理，每个卷积块两次卷积后进行池化。网络左侧的编码器由四个下采样操作构成，每个下采样部分包括两个 3×3 的卷积层和一个 2×2 的最大池化层，通过卷积提取图像中的上下文信息，然后进行池化降维，每下采样一次，特征图的尺寸即减半，编码器一共下采样 16 倍。网络的右侧为解码器部分，由四个上采样操作构成，其中每个上采样操作包括两个 3×3 的卷积层和一个上采样层，通过反卷积和上采样操作尽量还原与压缩部分同层图像大小，方便与对应编码部分的特征图进行简单跳跃连接，获得两倍大小的特征图组，再将其两次卷积得到重构特征，以此类推。

通过卷积获取图像的高级语义特征，对图像中感

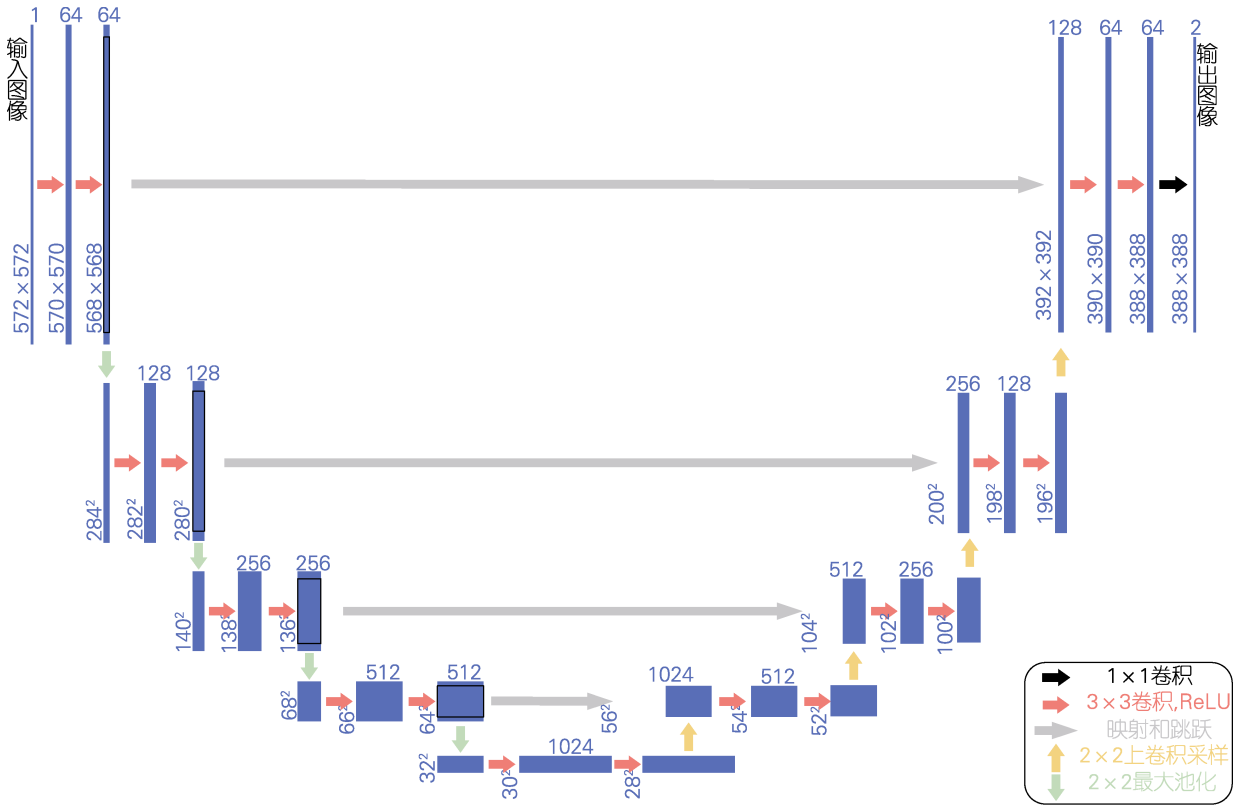

图 5　U-Net 网络结构

Fig.5 U-Net network structure

兴趣区域进行定位，并通过上采样层恢复卷积后特征图的分辨率。编码器和解码器通过跳跃连接融合网络浅层和深层的特征，在每个卷积层后连接非线性激活函数 ReLU 来增加网络的非线性特征，最后使用一个 1×1 的卷积层将 64 维的特征图映射为 2 维的输出，作为 U-Net 网络最终的分割结果图。

U-Net 各层输出见式（2）。

$$x_o=\begin{cases}Max_p(Conv_3(Conv_3(x_i))),1\leqslant L\leqslant 4\\ Conv_3(Conv_3(x_i)),L=5\\ Conv_3(Conv_3([U(x_i)])),6\leqslant L\leqslant 9\\ Conv_1(x_i),L=10\end{cases} \tag{2}$$

其中：x_i 和 x_o 为 U-Net 中第 L 层的输入和输出；$Max_p(\cdot)$ 为最大池化；$Conv_3(\cdot)$、$Conv_1(\cdot)$ 分别为 3×3 卷积操作和 1×1 卷积操作；$U(\cdot)$ 为上采样；[·]表示跳跃连接操作。

3 实验与结果分析

为验证论文所提方法有效性与可行性，本实验以文本生成图像的流程为例，演示虚拟人物形象设计和生成的过程和效果。

首先，选择用户喜爱度最高的“二次元风格”作为目标风格，并收集大量二次元风格图片，将其作为数据集（见图 6）对 Stable-Diffusion 网络进行参数微调；随后在此基础上进行虚拟人物形象新方案的设计与生成。

图 6 从网络中搜集的二次元风格图片[29-30]

Fig.6 Anime style pictures collected from the Internet[29-30]

其次，通过市场调查和用户分析，定义用户对目标形象的文本提示词如背景、人设、发色、衣服、人物姿态等。同时，在方案的设计与生成过程中，还需要加入反向提示词，用于约束生成网络输出结果，以避免生成的结果中出现与用户期望不符的内容。正向提示词和反向提示词，见表 1。由于文本生成图像的随机性很大，因此设置多批次输出图片，并从中初步筛选符合用户情感偏好的形象，筛选结果见图 7。

最后，将图像输入至图像生成图像网络中，再次对图片的细节和风格进行强化，继续设置为多批次输出图片，通过逐步迭代逐渐生成符合用户情感需求与偏好的虚拟形象，效果见图 8。

表 1 文本生成图像文本提示词与反向提示词

Tab.1 Text prompts and reverse prompts in text-generated image

正向提示词	意义	反向提示词	意义
highly detailed	生成结果包含很多细节	blurry	生成结果模糊
lively	生成人物有活泼的性格	watermark	生成结果中含有水印
smile	生成人物带有微笑	cropped	生成结果中出现裁剪
floating_hair	生成人物有飘逸的头发	missing fingers	生成结果中人物缺少手指
purple_eyes	生成人物有紫色的眼睛	worst quality	生成图像为最差的质量
…	…	…	…
Loli	生成人物的属性是萌妹	low quality	生成图像为低质量

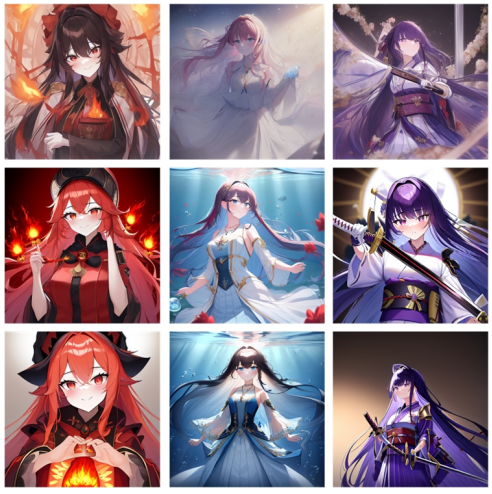

图 7 文本生成图像初步筛选结果

Fig.7 Preliminary screening results of text-generated image

图 8 图像生成图像迭代结果

Fig.8 Iteration result of image-generated image

通过图 7 和图 8 对比，可以看出生成形象经过图像生成图像的多次迭代及局部设计后，人物细节更加清晰、改善了手部、面部的畸形情况，图像构成变得和谐美观。

为更加有效表达用户情感，图像生成图像模型还可以对虚拟人物的表情进行修改，如图 9 所示，将按目标表情生成的原图（见图 9a）作为输入，设定虚拟人物的目标表情如开心、愤怒、冷漠等，可以得到对应的修改结果图（见图 9b—d），根据图 9 的结果可以看出，Stable-Diffusion 模型能够有效地基于设计的情感提示词修改虚拟人物的表情。

图 9　人物表情情感转换

Fig.9 Character expression and emotion conversion

4　结语

本文基于 Stable diffusion 模型研究了基于深度学习的虚拟形象设计和生成。并以本文生成图像的流程为例演示了设计和生成的过程和效果。相对于传统的设计方法，基于深度学习的方法有效地降低了设计门槛，不同于传统的专业人员设计、绘画，深度学习方法只需要收集相关风格的数据集，训练图像生成网络模型后便可以生成符合用户心理预期的各式各样的人物形象，并且其形象可随文本提示词变化，操作简便、生成快速的特点大大提高了虚拟人物设计的效率。同时也能够帮助用户更精准地实现符合其情感偏好的形象生成。Stable diffusion 模型具有图片生成稳定性强且质量高、生成速度快的特点，非常适合用于迭代生成和设计虚拟人物形象，但基于深度学习方法也存在以下不足与缺陷。

图像生成模型通常参数量非常庞大，因此针对某一风格图像生成还需要在该风格的数据集上进行参数微调才能得到正常的效果，如果输入模型没有训练过的风格图片，则无法输出正常结果；在生成人物形象的过程中，迭代中如果选到某些图片，可能会“收敛”到某一类型的图片，及时改变参数都很难发生变化，只能重新开始生成；从局部设计修改的结果可以看到如服装配色的大面积修改很容易引起其他内容变化，且在生成图像中面部和手部很容易出现畸形的效果。

基于深度学习方法的虚拟人物图像生成方法在应用上仍存在提升空间，未来将针对人物局部设计与人物风格调整方面展开重点研究与探索，以增强虚拟人物形象的灵活性和可塑性，从而适应不同用户的情感偏好与需求。

参考文献：

[1] 胡蓓蓓. 虚拟主播在音频生产和传播中的应用[J]. 广播电视信息, 2023, 30(3): 70-72.
HU Bei-bei. Application of Virtual Anchor in Audio Production and Communication[J]. Radio & Television Information, 2023, 30(3): 70-72.

[2] 郭嘉. 虚拟数字人在图书出版领域的多元身份构建研究[J]. 科技与出版, 2022(8): 56-63.
GUO Jia. Research on the Multi-Identity Construction of Virtual Digital Man in the Field of Book Publishing[J]. Science-Technology & Publication, 2022(8): 56-63.

[3] 喻国明, 耿晓梦. 试论人工智能时代虚拟偶像的技术赋能与拟象解构[J]. 上海交通大学学报(哲学社会科学版), 2020, 28(1): 23-30.
YU Guo-ming, GENG Xiao-meng. Technology Empowerment and Deconstruction of Virtual Idols in the Age of Artificial Intelligence[J]. Journal of Shanghai Jiao Tong University (Philosophy and Social Sciences), 2020, 28(1): 23-30.

[4] ROUCHITSAS A, ALM H. Ghost on the Windshield: Employing a Virtual Human Character to Communicate Pedestrian Acknowledgement and Vehicle Intention[J]. Information, 2022, 13(9): 420.

[5] 刘懿璇. “交互式沉浸”下文化社区虚拟形象的自我重构与社交体验[J]. 青年记者, 2022(24): 110-112.
LIU Yi-xuan. Self-Reconstruction and Social Experience of Virtual Image of Cultural Community under "Interactive Immersion"[J]. Youth Journalist, 2022(24): 110-112.

[6] 宋辰婷, 邱相奎. 数实交互与赛博集群者：虚拟形象直播中的认同建构[J]. 新视野, 2022(6): 54-61.
SONG Chen-ting, QIU Xiang-kui. Digital-Real Interaction and Cybercluster: Identity Construction in Virtual Image Live Broadcasting[J]. Expanding Horizons, 2022(6): 54-61.

[7] ZENG Jia-ying, ZHAO Ya-ling, YU Jing-yi, et al. Research on the Design of Virtual Image Spokesperson of

National Trendy Brand under the Concept of Meta-Universe[J]. Highlights in Art and Design, 2023, 2(2): 102-107.

[8] 裴文汐. 二次元文化背景下二维虚拟主播角色设计研究[D]. 无锡: 江南大学, 2022.
PEI Wen-xi. Research on the Role Design of Two- Dimensional Virtual Anchor Under the Background of Secondary Culture[D]. Wuxi: Jiangnan University, 2022.

[9] 牛旻, 陈刚. 基于虚拟偶像符号的品牌形象设计与传播[J]. 包装工程, 2021, 42(18): 282-286.
NIU Min, CHEN Gang. Brand Image Design and Communication Based on Virtual Idols[J]. Packaging Engineering, 2021, 42(18): 282-286.

[10] 谢新水. 虚拟数字人的进化历程及成长困境——以“双重宇宙”为场域的分析[J]. 南京社会科学, 2022(6): 77-87.
XIE Xin-shui. The Evolutionary History and Growth Dilemma of Virtual Digital Human: An Analysis of the Dual Universe Fields[J]. Nanjing Journal of Social Sciences, 2022(6): 77-87.

[11] 郝强, 郭子淳. 虚拟人物的身体图式分析[J]. 计算机辅助设计与图形学学报, 2020, 32(7): 1080-1086.
HAO Qiang, GUO Zi-chun. Analysis of Virtual Human in Body Schema[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(7): 1080-1086.

[12] ZHU Wen-min, FAN Xiu-min, ZHANG Yan-xin. Applications and Research Trends of Digital Human Models in the Manufacturing Industry[J]. Virtual Reality & Intelligent Hardware, 2019, 1(6): 558-579.

[13] 王晓红, 卢辉, 麻祥才. 基于生成对抗网络的风格化书法图像生成[J]. 包装工程, 2020, 41(11): 246-253.
WANG Xiao-hong, LU Hui, MA Xiang-cai. Generation of Stylized Calligraphic Image Based on Generative Adversarial Network[J]. Packaging Engineering, 2020, 41(11): 246-253.

[14] 张磬瀚, 孙刘杰, 王文举, 等. 基于生成对抗网络的文物图像修复与评价[J]. 包装工程, 2020, 41(17): 237-243.
ZHANG Qing-han, SUN Liu-jie, WANG Wen-ju, et al. Repair and Evaluation of Cultural Relic Images Based on Generative Adversarial Network[J]. Packaging Engineering, 2020, 41(17): 237-243.

[15] 王晓慧, 覃京燕, 全烘辰. 基于 AI 画作生成的个性化文化创意产品设计方法[J]. 包装工程, 2020, 41(6): 7-12.
WANG Xiao-hui, QIN Jing-yan, QUAN Hong-chen. Cultural and Creative Product Design Method Based on AI Painting[J]. Packaging Engineering, 2020, 41(6): 7- 12.

[16] MOKADY R, HERTZ A, ABERMAN K, et al. Null-Text Inversion for Editing Real Images Using Guided Diffusion Models[EB/OL]. (2023-01-14) [2023-03-06]. https://arxiv.org/abs/2211.09794.

[17] HAN I, YANG S, KWON T, et al. Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion[EB/OL]. (2023-01-14)[2023-03-06]. https://arxiv.org/abs/2303.08767.

[18] LEE S, HOOVER B, STROBELT H, et al. Diffusion Explainer: Visual Explanation for Text-to-Image Stable Diffusion[EB/OL]. (2023-01-14)[2023-03-06]. https://arxiv.org/abs/2305.03509.

[19] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision[EB/OL]. (2023-01-14)[2023-03-06]. https://arxiv. org/abs/2103.00020.

[20] LUO Huai-shao, JI Lei, ZHONG Ming, et al. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval and Captioning[J]. Neurocomputing, 2022, 508: 293-304.

[21] FU Jin-miao, XU Shao-yuan, LIU Hui-dong, et al. CMA-CLIP: Cross-Modality Attention Clip for Text-Image Classification[C]// 2022 IEEE International Conference on Image Processing (ICIP). Bordeaux: IEEE, 2022: 2846-2850.

[22] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding[EB/OL]. (2023-01-14) [2023-03-06]. https://arxiv.org/abs/2205.11487.

[23] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015: 234-241.

[24] SIDDIQUE N, PAHEDING S, ELKIN C P, et al. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications[J]. IEEE Access, 2021, 9: 82031-82057.

[25] LIU Zhen-qing, CAO Yi-wen, WANG Yi-ze, et al. Computer Vision-Based Concrete Crack Detection Using U-Net Fully Convolutional Networks[J]. Automation in Construction, 2019, 104: 129-139.

[26] SEVASTOPOLSKY A. Optic Disc and Cup Segmentation Methods for Glaucoma Detection with Modification of U-Net Convolutional Neural Network[J]. Pattern Recognition and Image Analysis, 2017, 27(3): 618-624.

[27] FALK T, MAI D, BENSCH R, et al. U-Net: Deep Learning for Cell Counting, Detection, and Morphometry[J]. Nature Methods, 2019, 16(1): 67-70.

[28] WANG Mo, WANG Jing, CUI Yun-peng, et al. Agricultural Field Boundary Delineation with Satellite Image Segmentation for High-Resolution Crop Mapping: A Case Study of Rice Paddy[J]. Agronomy, 2022, 12(10): 2342.

[29] anao_12. Genshin Impact Ganyu Pictures[EB/OL]. (2023-01-14)[2023-03-06]. https://cs14.pikabu.ru/post_img/big/2023/06/14/8/1686748836144887946.jpg.

[30] Brayan Osvaldo Morales Aréchiga (@Verius94). Genshin Impact Grass God Pictures[EB/OL]. (2023-01-14) [2023-03-06]. https://pbs.twimg.com/media/Fp5M3tWaEAEdzLj?format=jpg&name=large.

责任编辑：陈作