DOI:10.13878/j.cnki.jnuist.20240515002

邢润媚[1,2]　常升龙[3,4]　何宽[5,6]　朱曙光[5]　高琼[5]　胡昊[5,7]

AIGC图像质量评估指标研究

摘要

人工智能生成内容(AIGC)技术可为人类提供各种类型的信息生成服务,如何对AIGC进行准确的质量评估,是当前亟待解决的问题.本文主要针对大模型生成图像的质量及其评估指标开展深入研究.首先,从技术方面概述了当前评估AIGC的常见方法,如深度学习方法和计算机视觉方法等,介绍并分析了准确性、相关性、一致性、可解释性等指标在不同类型生成内容评估方面的表现.然后,为了展示评估指标的实际作用,以百度文心一言为例,对其生成的图像进行评估实验:使用直方图和噪点数量等量化指标对生成图像进行客观评估;使用整体协调性和美观性等视觉感官指标对生成图像进行主观评估.最后,综合对比客观评估和主观评估的结果,筛选出色偏、噪点数量、心理预期等AIGC产品质量评估的高可靠性指标.实验结果验证了综合使用主客观评估指标进行AIGC产品评估方法的有效性和可靠性.

关键词

人工智能生成内容;深度学习;计算机视觉;图像;质量评估

中图分类号 TP18

文献标志码 A

收稿日期 2024-05-15

资助项目 河南省高等教育教学改革研究与实践项目(2024SJGLX173,2019SJGLX690);河南省重点研发专项(231111210200,241111210300);中央引导地方科技发展专项(Z2022-1343001);黄河水利职业技术学院测绘地理信息职业教育研究课题(2021CHYB01)

作者简介

邢润媚,女,助教,研究方向为人工智能和大数据分析等.smile199103@163.com

胡昊(通信作者),男,教授,研究方向为智慧水利、水资源综合智能分析与调度等.85678199@qq.com

0　引言

AIGC即人工智能生成内容(Artificial Intelligence Generated Content),是指基于自然语言处理(Natural Language Processing,NLP)、机器学习(Machine Learning,ML)和深度学习(Deep Learning,DL)等技术,利用大模型框架自动或半自动地生成各种形式的文本、图像、音频和视频等多媒体内容[1-2].随着人工智能技术的飞速发展,AIGC技术已经开始为人类提供丰富多样的信息和服务,也成为相应领域的研究热点[3].AIGC技术可以分为基于规则的和基于机器学习的两大类.基于规则的AIGC技术,是指利用智能化专家系统,结合专业化知识库,通过编写规则的方法实现内容生成.其优点是可生成比较专业、准确的内容,缺点是编写规则的过程会耗费大量的人力、物力和时间.基于机器学习的AIGC技术,是指利用机器学习算法,通过学习、模拟成规模的数据以生成预期内容.其优点是可生成比较自然、流畅的内容,缺点是需要构建大规模语料库,且对计算资源的要求较高.

深度学习在图像识别、语音识别、自然语言处理、推荐系统等领域应用广泛,是近年来机器学习领域的研究热点.当前流行的AIGC技术是建立在深度学习基础之上的,深度学习为AIGC提供了理论、技术支撑和具体的实现方法[4],是助推AIGC应用爆发式增长的关键技术之一.随着深度学习技术的迅猛发展,AIGC的多场景应用正不断进行功能拓展和性能提升[5].

虽然兴起的时间不长,但AIGC已经拓展到了多个应用领域[6-7],如:1)新闻媒体,利用AIGC技术可以快速生成新闻稿件、摘要、标题等内容,大大提高了新闻媒体的效率和准确性;2)广告营销,利用AIGC技术可以快速生成广告文案、视频、图像等内容,帮助企业提高广告投放效率和转化率;3)电子商务,利用AIGC技术可以快速生成商品描述、评论和同类商品推荐等内容,帮助电商平台提高商品信息描述的丰富性和准确性,有效地增加用户的购买意愿;4)教育教学,

1　河南交通职业技术学院 物流学院,郑州,451460
2　郑州大学 土木工程学院,郑州,450001
3　河南师范大学 软件学院,新乡,453007
4　河南恒茂创远科技股份有限公司,郑州,450016
5　黄河水利职业技术学院 测绘工程学院,开封,475004
6　河南理工大学 测绘与国土信息工程学院,焦作,454000
7　华北水利水电大学 水利学院,郑州,450046

利用 AIGC 技术可以快速生成教学材料、辅助资料等内容,帮助教师或教育机构提高教学效率和质量.但不可忽视的是,AIGC 在实现其商业价值的同时也带来了一些挑战和风险[8],如:1)内容误导风险,AIGC 可能会被恶意利用以生成虚假、误导性信息,这将会对社会公共秩序和公共安全造成不良影响,甚至于引发社会事件;2)技术方面的挑战,AIGC 技术生成的内容可能存在语言不通顺、逻辑不清晰等问题,这将会影响其用户体验和商业化运营.因此,需要对基于 AIGC 技术及其生成内容进行标准化和规范化的探索,以客观评估和衡量 AIGC 产品的质量和效果.质量评估是确保 AIGC 可靠性和有效性的关键环节[9-10],而评估指标的选择对于衡量生成内容的质量至关重要.已有研究表明,AIGC 在多个不同领域具有可观的潜力[11],在一定程度上甚至具备代替人工的能力,且已经出现了相关的实例应用[12].因此,对在不同应用场景中使用 AIGC 技术产生的内容必须进行有效的量化评估,以指导用户正确自主有效使用 AIGC 产品,并帮助开发者升级符合实际需求的模型功能[13].

AIGC 的评估指标主要从质量、效率、创新、伦理等方面来考虑[14-15].1)质量:AIGC 是否符合人类普遍的审美标准,是否具有逻辑性、一致性和可信度,是否能够满足不同的场景和不同目标受众的需求和期望.2)效率:AIGC 的过程是否能够在较短的时间内完成,是否能够节省人力和物力资源(主要是硬件和算力),是否有潜力进一步提高生产内容的规模.3)创新:AIGC 是否具有独特性、新颖性和原创性,是否能够突破人类创作的局限和思维惯性,是否能够引发使用者的思考和灵感.4)伦理:AI 生成的内容是否符合社会的道德规范,是否尊重知识产权,是否能够避免虚假、误导和侵权等负面影响.

国际上对于 AIGC 的质量评估给予了高度关注[16-17],尤其是在文本领域,已经形成了一套相对完善的评估体系,具体如 ChatGPT、Google 内部使用的机器人聊天系统的评价等[18],包括准确性、相关性、一致性、可解释性等多个方面.虽然这些评估指标和方法已经应用于各种实际场景中[19-21],如不同场景下的文本和对话生成、图像识别、语音识别、视频推荐等,但都仅限于主观评价(美观度和内容协调性等),即人工评估,而在客观评价方面的研究相对不足.在 ChatGPT 等产品的应用浪潮下,国内众多学者、机构也对 AIGC 的质量评估进行了相关研究[6,22-23],但尚处于起步阶段.一些研究团队已经开始尝试使用大模型生成文本、图像、音频和视频等内容[24],并进行自定义指标下的质量评估.目前国内对于 AIGC 的评估主要集中在文本领域,而对于图像、音频和视频等形式的评估较少,且同样存在着主观评价指标较多,客观评价指标研究不足的问题.

随着技术的不断进步和创新应用场景的拓展,国内外对于多模态生成内容的需求在逐渐增加,对大模型产品生成内容的质量评估的需求也相应增加[25].本文旨在对使用 AIGC 技术生成的内容质量进行评估指标的探索,特别是针对大模型生成的文本、图像、音频和视频等内容形式的评估.为了具体展示评估指标的实际应用,本文使用百度的文心一言进行图像生成实验,并结合使用 Python 语言和 Open CV(Open Computer Vision)、PIL(Python Image Library)等图像质量分析工具包进行图像的直方图、失真度、噪点数量等客观指标的测量和对比.同时,还开展了主观实验,对 AI 生成图像的内容、细节、整体美观度和内容协调性等进行人工量化评估.通过客观和主观实验的综合对比,筛选出可靠度较高的评估指标.本研究可为 AIGC 生成产品的规范化和模型性能提升优化提供有价值的参考.

1 AIGC 原理及其质量评估

1.1 AI 生成图像的原理

AI 图像生成的实现主要是基于深度学习中的生成式对抗网络(GAN)等技术.深度学习技术通过卷积神经网络对大量图像的学习,使得模型能够自动地提取出图像中的特征和规律.GAN 是一种生成模型,通过训练两个神经网络(即生成器和判别器),来不断生成越来越逼真的图像.其中,生成器尝试生成假图像以欺骗判别器,而判别器则努力区分真实图像和假图像.两个神经网络互相竞争与合作,最终生成具有高度真实感的图像.

AI 生成图像的过程可以分为 3 个阶段:

1)训练阶段:模型通过学习大量的图像数据,从中提取出图像的特征和规律.

2)生成阶段:在训练阶段的基础上,模型根据一定的随机性生成新的图像.

3)优化阶段:通过 GAN 的优化调整,使得生成的新图像更加符合人们的视觉要求.

1.2 通用图像质量评估方法和指标

当前,对 AI 生成图像质量进行客观评估,可以

使用直方图、失真度、噪点数量等量化指标.直方图用于描述图像的亮度分布,失真度衡量的是生成图像的扭曲程度,噪点数量则反映图像中含有的随机噪声的数量.主观评估方面可以使用内容、细节、整体美观度和主要内容协调性等指标.主观指标主要通过人工评估来衡量图像的质量,如视觉上的清晰度、整体一致性和美观度等.目前还没有科学、统一的准则来评估 AI 生成图像的感知质量.为便于统计和对比,一般可从以下 5 个不同角度对生成的图像质量进行主客观评价[20].

1)技术问题(technical issues):可以理解为画面质量,如图像压缩情况等.

2)AI 伪影(AI artifacts):由于 AIGC 算法而导致的 AI 伪像.

3)不自然性(unnaturalness):违反常识的不自然现象和观看体验中的不适.

4)差异性(discrepancy):AIGC 生成的图像与期望之间的不匹配程度.

5)美学(aesthetics):AIGC 生成图像的整体视觉吸引力和美感.

评估生成图像的质量是一个复杂的任务,通常需要采用多种评估方法来综合判断生成图像的质量,并结合主观评价和客观指标,以获得更全面的理解.

1.3 图像质量评估与人眼视觉的关系

探讨大模型生成图像的质量评估方法和指标,不能脱离用户的使用感受.而用户对大模型性能的评估是通过人眼目视,将视觉反应与心理预期进行对比,获得直接的感受.无论是生成图像还是通过各种传感器获得的图像,其最终形式多为电子化的图像,即具备灰度值的像素组合,这与人眼分辨物体的感官机能是一致的[26].所以,(电子)图像与人眼视觉之间存在密切的关系,在图像处理和显示技术中,需要考虑人眼的视觉特性,以提供更加逼真、舒适的视觉体验[27].

一般从心理物理量(亮度、主波长和纯度)和相应的心理量(明度、色度和饱和度)两个维度来探讨视觉的核心特性[28].亮度作为描述光的强度的参数,它与物体表面或光源的实际亮度密切相关.但物体或光源的实际亮度高并不一定导致人感知到的明度也高.光谱是由多种不同波长的光组合而成的,而不同波长所引发的视觉感知便是色度,色度反映了不同波长光给人的颜色感觉.纯色描述的是那些未混入白色成分的窄带单色光,它在视觉上呈现为高饱和度的颜色.实际上,可见光谱中的各种单色光都是最为饱和的彩色,但当光谱色中混入的白光成分增多时,其饱和度会随之降低,表现为颜色的不饱和.

基于上述图像与人眼视觉的关系,本研究设计了 AI 生成图像的质量客观评估实验和主观评估实验,以探索更加符合人眼视觉与心理预期的评估参数.

1.4 AI 生成图像评估实验方案

1.4.1 客观评估实验方案

基于当前 AIGC 算法在各领域的应用成熟度情况,以及国内的中英文使用情况[29],本文选择使用百度文心一言进行图像生成相关实验.文心一言是百度研发的新一代知识增强大语言模型,能够与人对话互动、回答问题、协助创作,高效便捷地帮助人们获取信息、知识和灵感.文心一言融合了数万亿数据和数千亿知识点,并学习得到预训练大模型.作为扎根于中文市场的大语言模型,文心一言具备中文领域最先进的自然语言处理能力.对比国际上领先的 ChatGPT、Midjourney 等分别在文本和图像生成方面公认领先的大模型,文心一言在中文语言和中国文化上有更好的表现.

为了更好地测试文心一言“以图生图”功能生成图像的质量,本研究使用了 Lenna 图作为原图.作为长期以来业内最流行的标准测试图,Lenna 图包含了平坦区域、阴影和纹理等细节,符合测试的特殊要求.例如,处于低频区域的光滑皮肤、镜面,处于高频区域的羽毛、繁杂饰物等,可用于测试各种不同的图像处理算法.本研究中使用南加州大学网站获取的 Lenna 图原始扫描电子版,并将其输入到文心一言作为参考真值,使用相应的插件生成“画质修复图”、“AI 重绘图”、“相似图”等.其中的插件包括“AI 重绘”、“画质修复”和“生成相似图”等.为了确保使用 AIGC 生成的图像符合实际,研究人员参考了文心一言上使用频度较高的关键热词,并根据提示词(指令)有目标地生成了约 10 个类别的 1 000 张热门图像.相应的图像生成功能及其内在算法可以通过 Python 编程语言中的图像处理库来实现.例如,Python 中的 PIL 和 Open CV 工具库提供了各种图像处理函数和算法,可以方便地使用这些函数来自动化评估图像质量.本研究中的客观评估实验将彩色直方图、清晰度、亮度、色偏、噪点数量统计和失真程度等客观指标结合起来,使用多种算法进行综合评估,以计

算图像的总体质量得分,获得更全面、客观的评估结果.

1)彩色直方图:直方图是从图像内部灰度级的角度对图像进行表述的,统计的是图像内各灰度级出现的次数.通过直方图可以清晰地观察到图像的整体灰度分布,便于图像的后续分析和处理.获得图像的直方图就是统计灰度级(即像素值)出现的总频数的过程,其计算公式[30]如下:

$$h(r_k)=n_k,\quad k=0,1,2,\cdots,L-1. \tag{1}$$

其中:r_k 为像素的灰度级;n_k 是具有灰度 r_k 的像素个数.

通常将灰度级出现的总频数除以总像素数,以概率的形式表述彩色直方图的概念:

$$p(r_k)=\frac{h(r_k)}{N_t}=\frac{n_k}{N_t}. \tag{2}$$

其中:N_t 为总像素数.

2)清晰度:指利用拉普拉斯算子或者 Sobel 算子[31]计算图像的二阶导数,反映了图像的边缘信息.清晰度高,则相应的方差值就更大.用于点阵数码影像时,其单位为 DPI(Dots Per Inch),表示图像每英寸长度内的像素点数,即指每一英寸长度中,取样、可显示或输出点的数目.

3)亮度:指计算图像在灰度图上的均值和方差.当存在亮度异常时,均值会偏离均值点,相应的方差值也会偏小,据此可评估图像是否存在过曝光或曝光不足的问题[31].简单来说,亮度即一幅图像给观察者的一种直观感受.如果图像是灰度图像,则其亮度与其灰度值有关,灰度值越高则图像越亮.

4)色偏:又叫色差,是指拍摄的图像中某种颜色的色相、饱和度与真实的图像有明显的区别,而这种区别通常不是人们所希望的.将 RGB 图像转变到 CIE Lab 空间进行分析,其中,L 表示图像亮度,a 表示图像红/绿分量,b 表示图像黄/蓝分量.存在色偏的图像,在 a 和 b 分量上的均值一般会较多地偏离原点,相应的方差值偏小.因此,可通过计算图像在 a 和 b 分量上的均值和方差来评估图像是否存在色偏.色差 $\Delta E_{ab}=1$ 时称为 1 个 NBS(美国国家标准局)色差单位,1 个 NBS 单位大约相当于视觉色差识别阈值(颜色宽容度)的 5 倍[32].

5)噪点数量:使用噪点检测算法来计算图像中的噪点数量.较常用的噪点检测算法是高斯滤波器[33],可以将图像中的噪点模糊化,然后通过计算处理前后的差异统计噪点数量.噪点数量的单位即统计出所有被认为是噪声的像素个数.

6)失真程度:使用失真评估算法来计算图像的失真程度.较常用的失真评估算法是均方误差(MSE)算法[34],它通过计算图像中每个像素与标准图像像素之间的差异,对其求平方和,并取平均值,从而计算整个图像的失真程度.本研究中使用峰值信噪比(Peak Signal to Noise Ratio,PSNR)表示失真程度,单位为 dB,PSNR 值越大表示失真度越小.

在以上客观评估实验中,各个图像信息的自动化提取及其评估指标的计算,将综合使用 Python、Open CV、Matplotlib、PIL 等技术或工具具体实现.

1.4.2 主观评估实验方案

随机挑选若干研究人员作为实验参与人员(测试者),对 AI 生成图像进行质量评估(主观评分).为最小化外部条件所造成的差异,所有测试者主观实验测试条件均相同,并遵循主观测试建议(ITU-R BT.500-13,即国际电联无线电通信部门(ITU-R)电视图像质量的主观评价方法)[33].测试者在同一个具有正常室内照明的实验室环境中,坐在距离电脑屏幕大约 1.5 倍屏幕高度(45 cm)的位置.具体做法为:随机选取 20 人作为测试者(第一作者所在学校的学生),包括 10 名男性和 10 名女性,要求他(她)们根据指定的评价方法和准则对文心一言生成的图像进行质量评估(评分),评分参考指标有提示词匹配度、整体协调度和整体美观度等 6 项指标.此 20 名测试者只进行每张图像的评分,不参与图像生成或评分统计等其他活动.每张图像的最终得分为 20 名测试者给出分数的均分.图像评分的计算公式为

$$S=\frac{t+s_c+s_t}{O}+\frac{W_1c+W_2a+W_3p}{3}. \tag{3}$$

其中:$O=6$,是指标总数量;t 是类型评分,s_c 是场景评分,s_t 是风格评分,这三项均属于“是否与提示词匹配”的子栏目;c 是整体协调度;a 是整体美观度;p 为是否达到心理预期;W_i 是根据相应指标的重要性而赋予的权重,本例中,$W_1=0.1$,$W_2=0.2$,$W_3=0.3$[34].

2 实验验证

2.1 客观实验结果及图像质量评估

为了便于量化对比,首先制作 Lenna 原图和 AIGC 生成图的各自彩色直方图.实验结果如图 1 所示.

除了进行直方图的对比,本文还基于 Open CV

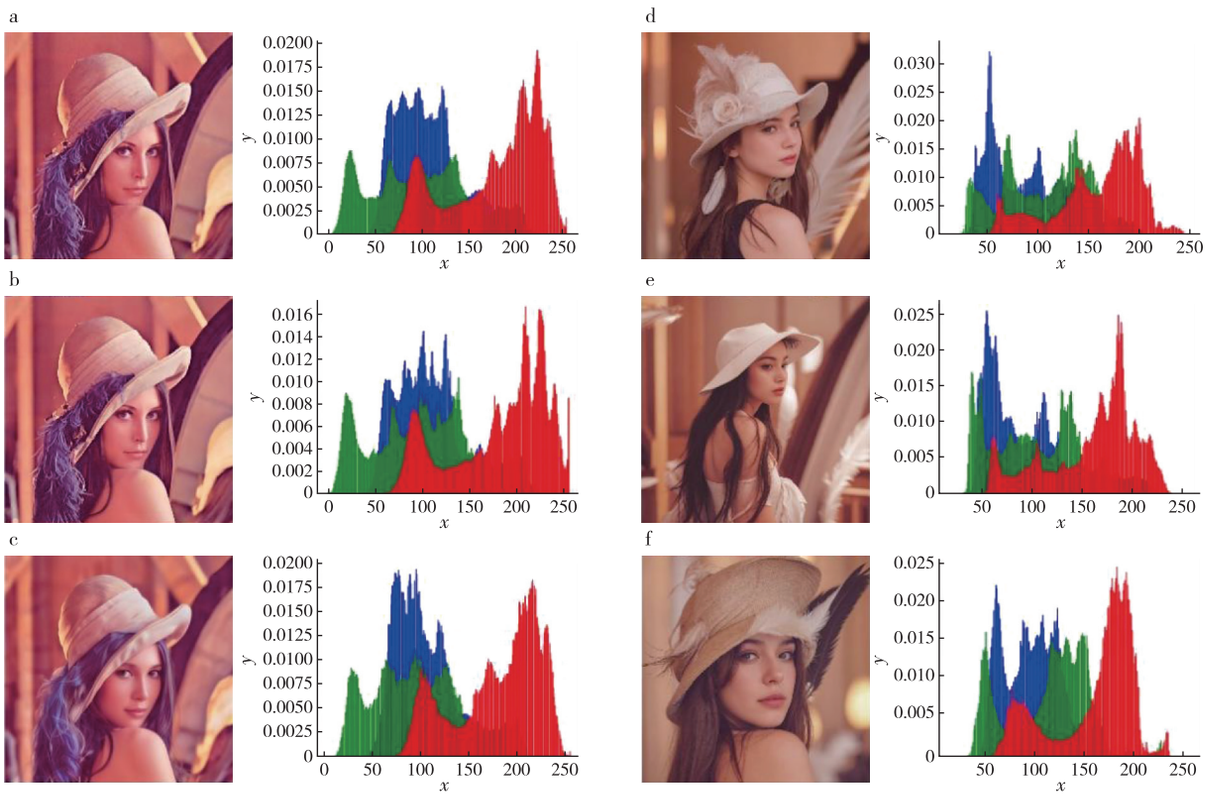


图 1　图像生成结果及各自图像的 RGB 通道信息(每张图的右侧为各自的彩色直方图;横坐标是灰度级,纵坐标是灰度级出现的频率)　a.Lenna 图原图;b.画质修复生成图;c.AI 重绘生成图;d-f.根据 Lenna 原图使用“以图生图”功能生成的系列相似图

Fig. 1　Image generation results and corresponding RGB channel information(the horizontal axis represents grayscale level,and the vertical axis represents the frequency of occurrence of each grayscale level)
a.original Lenna image;b.restored image with enhanced quality;c.AI-redrawn image;d-f.a series of similar images using the image-to-image function based on original Lenna image.On the right side of each image are their respective colored histograms

技术对 Lenna 原图及各个 AI 生成图进行了评估.评估指标包括清晰度、亮度(因灰度值没有单位,此处将其归一化为 0~1 间的值)、色偏、噪点数量统计和失真程度等客观指标的对比,结果如表 1 所示.

2.2　主观实验结果及图像质量评估

对 AI 生成图像的主观评价实验同样使用文心一言:首先,进行图像生成实验;然后,随机选定学生 20 人,参与对 AI 生成图像的主观实验及图像质量评价.生成图像中具有代表性的图像如图 2 所示.

AI 图像所使用的提示词(prompts)不同,生成的主题和内容则不同.图 2 中,不同行的图使用的是不同的提示词,同一行的子图使用的则是同样的提示词,即用同样的提示词反复生成大量同一主题图像,然后选择其中具有代表性的图像进行对比和评估.

表 1　原图与生成图的图像质量指标对比

Table 1　Comparison of image quality indicators between original image and generated images

生成图类型	清晰度/DPI	亮度(灰度值)	色偏(ΔE)	噪点数量/个	失真程度/dB
Lenna 原图(图 1a)	38.243	0.097	2.487	227 514	107.675
画质修复图(图 1b)	99.545	0.068	**2.537**	**192 892**	109.220
AI 重绘图(图 1c)	116.187	0.131	2.577	57 482	**107.859**
相似图(图 1d)	**36.413**	0.171	2.900	161 831	106.974
相似图(图 1e)	27.663	0.166	2.598	158 625	106.110
相似图(图 1f)	99.400	**0.118**	2.689	110 071	107.142

注:加粗字体表示最优值.

图 2　根据不同的提示词生成的代表性图像

Fig. 2　Representative images generated based on different prompt words

图 2 中生成各类型子图所使用的提示词(提示词为生成该类型图下的某一张图像所使用的所有输入文字内容(对话))如表 2 所示.

根据前述主观评价实验方案进行 AI 生成图像(图 2)的主观评价,计算结果如表 3 所示.各个指标均归一化为 0~1 的值,值的大小反映相应的图像给测试者的不同主观感受,如:“场景”对应的值 1.0,表示完全与提示词匹配,而 0.3 则表示与提示词匹配程度较低;“心理预期度”的值 0.8,表示 AI 生成的图像整体比较符合测试者的心理预期,而 0.5 则表示测试者对于 AI 生成图像相对不是很满意,不太符合心理预期.“总分”指标在所有细分主观指标的基础上,通过客观计算方法获得(式(3)),一定程度上避免了绝对主观评分的偶然性和不稳定性,能够更加准确地反映测试者对 AI 生成图像的主观评价.

3　结果讨论

3.1　对客观实验结果的分析和讨论

本研究的目的是探索关于 AIGC 质量的评估指标,并以文心一言的绘图功能和插件进行图像生成实验,以筛选验证合适的指标.对比观察客观实验结果发现:从直方图显示的信息来看,图 1 中 AI 重绘和画质修复效果几乎相同,但 AI 重绘的 RGB 三通道各自值的分布呈现更明显的聚类效果.AI 重绘图与原图差别更大,而画质修复图呈现与原图更类似的自然真实性.相比 AI 重绘和画质修复图,三幅相似图与原图的差别更大,这也可以通过直方图中 R、G、B 值的分布区间差异观察出来.因此,直方图在 AIGC 生成图质量评估中具有一定的指示作用.

目视对比观察三幅相似图,可看出图 1e 虽然风格、结构等与另外两张图相似,但图中人物形态扭曲

表 2　生成图像所用的提示词

Table 2　Prompt words used to generate images

序号	类型	场景	风格	提示词
图 2a	人物图画	办公室	漫画	生成一幅人物图画,漫画风格,内容是办公室工作的职员
图 2b	风景画	海边	梵高绘画	生成一幅风景画,梵高绘画的风格,内容是海边
图 2c	人物图画	名画仿制	水彩画	生成一幅人物图画,水彩画风格,内容是达·芬奇的名画《蒙娜丽莎》
图 2d	人物图画	篮球场	写实	生成一幅人物图画,写实摄影风格,主题内容是篮球场上的运动员在奋力拼抢
图 2e	动物图画	草地	写实	生成一幅小猫的图画,写实风格,内容是小猫在草地上玩

表 3　对生成图像的主观评价结果

Table 3　Subjective evaluation of generated images

序号	是否与提示词匹配			协调度	美观度	心理预期度	总分
	类型	场景	风格				
图 2a	1.0	1.0	1.0	0.6	0.6	0.7	0.63
图 2b	1.0	1.0	0.5	0.8	0.9	0.8	0.58
图 2c	0.5	0.3	0.6	0.7	0.9	0.6	0.38
图 2d	1.0	0.9	1.0	0.3	0.6	0.5	0.58
图 2e	1.0	1.0	0.5	0.9	0.9	0.7	0.58

(胳膊与身体的构造产生错误),而这种错误不能体现在直方图上.观察表 1,进一步对比原图与 AIGC 生成图在各个客观指标上的量化差异.相较于直方图,从表 1 中可直接计算出在各个指标指示下与原图最接近的值(即表 1 中加粗数值).其中,画质修复图产生了两个最佳值(分别是色偏值与噪点数量),AI 重绘图产生了一个失真程度最佳值,生成的相似图 1d 与相似图 1f 分别在清晰度和亮度两个指标上与原图最接近.只有相似图 1e 没有产生任何一个最佳指标值,这与人眼目测观察所得一致,即画质修复图的效果与原图最近似,而生成的相似图 1e 与原图差异最大.由此可见,色偏与噪点数量两个指标的指示作用较为准确.

3.2　对主观实验结果的分析和讨论

因为需要从用户的角度找出模型输出结果的错误,以便更全面地评测模型的性能,所以图 2 中选用的部分图像,并非都是具备完美效果的绘制图.下面对生成图像中典型的、共性的错误进行分析和讨论(图 2).

如图 3 中各子图所示:

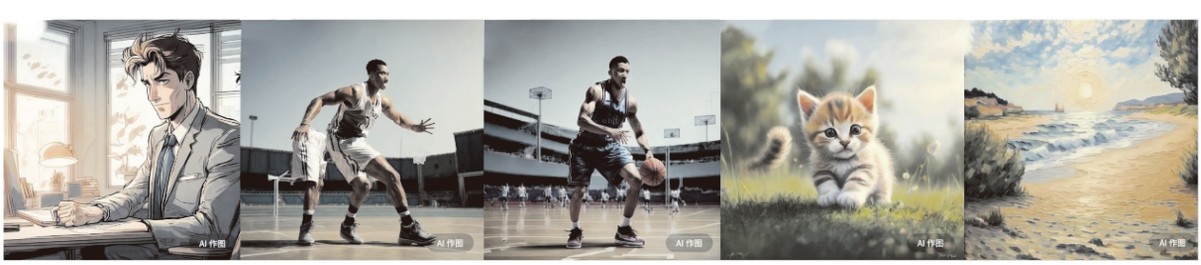
1)人物图的手部不自然或明显错误."画手指难"是 AI 绘画领域长期以来难以解决的问题,国外先进的 Midjourney 模型和国内的众多大模型均是如此.如图 3a 中人物的手部有很明显的错误,右手的手指多,左胳膊完美但没有手部.其他生成图中这种现象也存在,即使提示词中的主题并非以人物为主要生成任务.

2)脸部扭曲、五官错位.生成图 3b 的提示词中有"人物图画"的字眼,但是并没有对五官有局部特写或细节要求,图像整体的写实摄影风格完美匹配,但是五官明显变形,非常不自然.且尽管提示词没有"手"的字眼,模型还是绘制了有明显错误的手部.

3)肢体错位.生成图 3c 的提示词与图 3b 一样,是同一个主题,提示词中有"人物图画"的字眼,并没有对人物的肢体尤其下肢有特写或细节要求,却出现了非常明显的肢体错位.AI 生成的图像极难完善眼睛、手、脚等部位细节,推测之一是因为神经网络没有足够的数据学习手指与手指之间的结构逻辑,且手指关节间的特征属于细小颗粒[10],因此生成的手容易出错.与手部不同,人物的下肢肢体并非属于细小颗粒,但生成的结果同样出现较明显的错误,除了提示词未给出明确提示外(如"该人有两条腿"),原因可能是大模型的生成模式和方法仍然处于一种"懵懂"的混沌状态,类似人类幼童的学习阶段.这与网络模型的"黑盒"模式一样,暂时无法解释或推测其原理,只能是有目标地训练模型,其做法类似于机器学习中的监督学习.

由图 3 可知,"不会画手"是文心一言的突出问题,其实这也是很多大模型的通病.因此,图像中的手是否有错误,也成为判断一幅人物图是否为 AI 作图的一个重要标识.目前的大模型中,除了 MidjourneyV5 勉强能达到预期外,其他 AI 作图大模型生成图像中眼睛、手、脚等部位的细节存在明显问题.

4)动物图像绘制细节问题.图 3d 的整体风格,尤其是猫的毛、五官和草地等细节处理非常完美,但是猫爪出现了与人手一样的问题,而且猫的尾巴与身体分离,这与人物图肢体错位类似.据此分析,大



a. 手部变形或不自然　b. 脸部五官变形或不自然　c. 肢体错位　d. 动物图中绘制细节问题　e. 生成图的内容空间分布逻辑问题

图 3　人物生成图中内容的共性错误(a-c)以及非人物生成图中内容的共性错误(d-e)

Fig. 3　Common errors in generated content,(a,b,c) for human-themed images,(d,e) for non-human-themed images

模型之所以出现这种现象,很可能是因为在学习过程中接触了大量侧面视角的人手或动物爪子图像,而这些图像未能完整展示出人手或猫爪的具体手指数量.至于肢体错位问题,推测是由于大模型过多地学习了非正常视角(如遮挡或侧视图)的图像,导致在特征提取与空间位置、逻辑关系之间的匹配度不足,从而引发了生成图像的错乱.然而,鉴于大模型在处理更为精细的图像细节时的出色表现(如毛发、草地等),可以排除是由于颗粒细度不足导致大模型无法正确绘制人手或五官的假设.

5)内容不合乎空间分布逻辑.图 3e 整体风格协调、美观,但是左侧的海浪与沙滩连为一体,而非像右侧一样自然而分明,且在这样狭小的海湾中海浪的高度太高,不符合常识.虽然在提示词中选用了梵高绘画的风格,但该景观显然并非大自然中真实存在的景象.由此可见,图像中的空间和逻辑问题对于大模型来说似乎是无法有效学习和理解的难题,所以会生成带有空间分布逻辑错误的图像.

通过实验验证,上述错误图像的生成不能仅归因为文本生成中常见的语义(提示词)理解偏差问题,因为同样的提示词重复生成图像时,也会产生很多没有任何问题的“完美”图像.目前所见的大模型,其核心代码或语料库或图像训练库未完全开源,大多研究者对于模型训练机制不得而知.但从结果来看,各大模型产品的核心仍然是以卷积神经网络为主干的深度学习模型.而且训练数据库中的样本图像是人工标注,仍然属于深度学习对图像处理工作中的图像分类和语义描述范畴.即使在中文领域表现良好的文心一言,其使用的仍然是英文标注的训练数据.这是因为当前深度学习领域中的公开数据集大多为欧美机构开发,其使用的标注语言为英文.通过上述讨论与表 3 的量化结果,可发现某些主观指标与最终图像质量的总体评分是一致的,特别是在美观度、协调度和心理预期匹配度等方面.这些指标在评估中表现出了较高的可靠性,对最终图像质量的评估起到了有效的指示作用.

本研究通过客观实验对比了直方图、失真度和噪点数量等客观指标的效果,筛选出色偏和噪点数量两个指示性相对准确的指标.在主观实验中,评估了图像的内容、细节、整体美观度和主要内容协调性等指标,发现文心一言生成的图像得到较高的评价.

大模型生成图像最常见的问题是生成图像的空间和逻辑错误较多,这可能是由于模型在生成图像时过度强调细节和复杂度,导致生成的图像与真实场景的差异较大[35].此外,大模型生成图像还存在内容的协调性和风格两者不一致的问题,这可能是因为模型在标注数据集标签时未充分考虑到内容和风格的一致性.在实际应用中,逻辑性、内容协调性等方面的错误有可能对用户造成误导[36].对于空间和逻辑错误较多的问题,可尝试优化模型的生成算法,减少对细节和复杂度的敏感性和过度追求,以更多地关注图像的整体质量和清晰度;对于内容和风格不协调、不一致的问题,可尝试引入包含更多的语义信息和风格特征的数据集,并加强标注的准确性,使模型在生成图像时能够更好地理解和把握内容的整体风格和特点.

4 结论

本研究介绍了大模型图像生成质量的评估方法,对比分析了准确性、相关性、一致性、可解释性、直方图、失真度、噪点数量等客观指标,以及内容、细节、整体美观度、内容协调性和心理预期匹配度等主观指标,还使用百度文心一言的图像生成功能进行了主观实验,并进行人工评估.通过对比主客观实验结果,发现客观指标如色偏和噪点数量,主观指标如心理预期等能够有效地评估图像质量.美观度、协调度和心理预期匹配度等指标,在评估中表现出较高的可靠性.因此,对用户使用提示词的掌握程度和作品完成度进行评价,可帮助用户在各种使用场景下做出更准确的评估.未来可在深入理解 AIGC 基础上,设计更合理的评价指标,探索更有效的评估方法,以提高生成内容的整体质量评估水平.

参考文献
References

[1] 李白杨,白云,詹希旎,等.人工智能生成内容(AIGC)的技术特征与形态演进[J].图书情报知识,2023,40(1):66-74
LI Baiyang, BAI Yun, ZHAN Xini, et al. The technical features and aromorphosis of artificial intelligence generated content (AIGC) [J]. Documentation, Information & Knowledge, 2023, 40(1): 66-74

[2] Wen J B, Kang J W, Xu M R, et al. Freshness-aware incentive mechanism for mobile AI-generated content (AIGC) networks [C]//2023 IEEE/CIC International Conference on Communications in China (ICCC). August 10-12, Dalian, China. IEEE, 2023: 1-6

[3] 朱永新,杨帆.ChatGPT/生成式人工智能与教育创新:机遇、挑战以及未来[J].华东师范大学学报(教育科

学版),2023,41(7):1-14
ZHU Yongxin, YANG Fan. ChatGPT/AIGC and educational innovation: opportunities, challenges, and the future [J].Journal of East China Normal University (Educational Sciences),2023,41(7):1-14

[4] Cao Y H,Li S Y,Liu Y X,et al.A comprehensive survey of AI-generated content (AIGC):a history of generative AI from GAN to ChatGPT [J]. arXiv e-Print, 2023, arXiv:2303. 04226

[5] 曲艺,刘海燕,曹玉东.基于多尺度卷积神经网络的无参考图像质量评价[J].辽宁工业大学学报(自然科学版),2024,44(2):115-120
QU Yi, LIU Haiyan, CAO Yudong. Non-reference image quality evaluation based on multi-scale convolutional neural network [J]. Journal of Liaoning University of Technology (Natural Science Edition), 2024, 44(2): 115-120

[6] 陈向东,褚乐阳,王浩,等.教育数字化转型的技术预见:基于 AIGC 的行动框架[J].远程教育杂志,2023,41(2):13-24
CHEN Xiangdong,CHU Leyang,WANG Hao,et al.Technology foresight in digital transformation of education:action framework based on AIGC[J].Journal of Distance Education,2023,41(2):13-24

[7] 王常圣.人工智能驱动的数字图像艺术创作:方法与案例分析[J].智能科学与技术学报,2023,5(3):406-414
WANG Changsheng.AI-driven digital image art creation: methods and case analysis[J].Chinese Journal of Intelligent Science and Technology,2023,5(3):406-414

[8] 李亚玲,覃缘琪,魏阙.人工智能生成内容的潜在风险及治理对策[J].智能科学与技术学报,2023,5(3):415-423
LI Yaling, QIN Yuanqi, WEI Que. Potential risks and governance strategies of artificial intelligence generated content technology[J].Chinese Journal of Intelligent Science and Technology,2023,5(3):415-423

[9] 宋士杰,赵宇翔,朱庆华.从 ELIZA 到 ChatGPT:人智交互体验中的 AI 生成内容(AIGC)可信度评价[J].情报资料工作,2023,44(4):35-42
SONG Shijie, ZHAO Yuxiang, ZHU Qinghua. From ELIZA to ChatGPT:AI-generated content (AIGC) credibility evaluation in human-intelligent interactive experience[J].Information and Documentation Services, 2023,44(4):35-42

[10] 吴柯烨,孙建军,谢紫悦.基于专利文本挖掘的细粒度技术机会分析[J].情报学报,2023,42(10):1199-1212
WU Keye, SUN Jianjun, XIE Ziyue. Research on fine-grained technology opportunity analysis based on patent text mining[J].Journal of the China Society for Scientific and Technical Information,2023,42(10):1199-1212

[11] 毕文轩.生成式人工智能的风险规制困境及其化解:以 ChatGPT 的规制为视角[J].比较法研究,2023(3):155-172
BI Wenxuan.The dilemma in the risk regulation of generative artificial intelligence and its resolution:taking ChatGPT as an example[J]. Journal of Comparative Law, 2023(3):155-172

[12] 宋一飞,张炜,陈智能,等.数字说话人视频生成综述[J].计算机辅助设计与图形学学报,2023,35(10):1457-1468
SONG Yifei,ZHANG Wei,CHEN Zhineng,et al.A survey on talking head generation [J]. Journal of Computer-Aided Design & Computer Graphics, 2023, 35(10): 1457-1468

[13] 林懿伦,戴星原,李力,等.人工智能研究的新前线:生成式对抗网络[J].自动化学报,2018,44(5):775-792
LIN Yilun,DAI Xingyuan,LI Li,et al.The new frontier of AI research:generative adversarial networks[J].Acta Automatica Sinica,2018,44(5):775-792

[14] 汪波,牛朝文.从 ChatGPT 到 GovGPT:生成式人工智能驱动的政务服务生态系统构建[J].电子政务,2023(9):25-38
WANG Bo,NIU Chaowen.From ChatGPT to GovGPT:the construction of government service ecosystem driven by generative artificial intelligence[J].E-Government,2023(9):25-38

[15] 严昊,刘禹良,金连文,等.类 ChatGPT 大模型发展、应用和前景[J].中国图象图形学报,2023,28(9):2749-2762
YAN Hao,LIU Yuliang,JIN Lianwen,et al.The development,application,and future of LLM similar to ChatGPT [J]. Journal of Image and Graphics, 2023, 28(9): 2749-2762

[16] Wu F, Hsiao S W, Lu P. An AIGC-empowered methodology to product color matching design[J]. Displays,2024,81:102623

[17] Liu G Y,Du H Y,Niyato D,et al.Semantic communications for artificial intelligence generated content (AIGC) toward effective content creation [J]. arXiv e-Prints, 2023,arXiv:2308. 04942

[18] 陈兵,董思琰.生成式人工智能的算法风险及治理基点[J].学习与实践,2023(10):22-31
CHEN Bing, DONG Siyan. Algorithm risks and governance bases of generative artificial intelligence[J]. Study and Practice,2023(10):22-31

[19] Li C,Zhang C,Waghwase A,et al.Generative AI meets 3D:a survey on text-to-3D in AIGC era [J]. arXiv e-Print,2023,arXiv:2305. 06131

[20] Zhang Z C,Li C Y,Sun W,et al.A perceptual quality assessment exploration for AIGC images[C]//2023 IEEE International Conference on Multimedia and Expo (ICME).July 10-14,2023,Brisbane,Australia.IEEE,2023: 440-445

[21] Wang T,Zhang Y S,Qi S R,et al.Security and privacy on generative data in AIGC: a survey [J]. arXiv e-Print, 2023,arXiv:2309. 09435

[22] 王静静,叶鹰.生成式 AI 及其 GPT 类技术应用对信息管理与传播的变革探析[J].中国图书馆学报,2023,49(6):41-50

WANG Jingjing,YE Ying.A probe into the generative AI and GPT-type technical applications with transform for information management and communication[J].Journal of Library Science in China,2023,49(6):41-50

[23] 王华树,刘世界.智慧翻译教育研究:理念、路径与趋势[J].上海翻译,2023(3):47-51,95
WANG Huashu,LIU Shijie.Smart translation education:concept,pathways and prospects[J].Shanghai Journal of Translators,2023(3):47-51,95

[24] 万小军.智能文本生成:进展与挑战[J].大数据,2023,9(2):99-109
WAN Xiaojun.Intelligent text generation:recent advances and challenges[J].Big Data Research,2023,9(2):99-109

[25] 祝智庭,戴岭,胡姣.高意识生成式学习:AIGC 技术赋能的学习范式创新[J].电化教育研究,2023,44(6):5-14
ZHU Zhiting,DAI Ling,HU Jiao.Higher consciousness generative learning:innovation of learning paradigm enabled by AIGC technology[J].e-Education Research,2023,44(6):5-14

[26] 张熙,杨小汕,徐常胜.ChatGPT 及生成式人工智能现状及未来发展方向[J].中国科学基金,2023,37(5):743-750
ZHANG Xi,YANG Xiaoshan,XU Changsheng.Current state and future development directions of ChatGPT and generative artificial intelligence[J].Bulletin of National Natural Science Foundation of China,2023,37(5):743-750

[27] 于天河,柳梦瑶.基于人眼视觉系统的图像质量评价方法[J].北京邮电大学学报,2023,46(2):129-136
YU Tianhe,LIU Mengyao.Image quality evaluation method based on human visual system[J].Journal of Beijing University of Posts and Telecommunications,2023,46(2):129-136

[28] 柳梦瑶.基于人眼视觉系统的图像质量评价方法研究[D].哈尔滨:哈尔滨理工大学,2022
LIU Mengyao.Research on image quality evaluation method based on human visual system[D].Harbin:Harbin University of Science and Technology,2022

[29] Lu Z Y,Huang D,Bai L,et al.Seeing is not always believing:a quantitative study on human perception of AI-generated images[J].arXiv e-Print,2023,arXiv:2304.13023

[30] Hassan M,Bhagvati C.Structural similarity measure for color images[J].International Journal of Computer Applications,2012,43(14):7-12

[31] 张彦超.基于边缘和颜色特征的图像检索技术研究[D].武汉:武汉理工大学,2010
ZHANG Yanchao.The research of image retrieval based on edge and color feature[D].Wuhan:Wuhan University of Technology,2010

[32] 杨杨.基于均匀色差空间扩展的彩色图像质量评价研究[D].合肥:中国科学技术大学,2013
YANG Yang.Research of color image quality assessment based on expanded uniform color difference space[D].Hefei:University of Science and Technology of China,2013

[33] 谢勤岚.图像降噪的自适应高斯平滑滤波器[J].计算机工程与应用,2009,45(16):182-184
XIE Qinlan.Adaptive Gaussian smoothing filter for image denoising[J].Computer Engineering and Applications,2009,45(16):182-184

[34] 魏政刚,袁杰辉,蔡元龙.一种基于视觉感知的图像质量评价方法[J].电子学报,1999,27(4):79-82
WEI Zhenggang,YUAN Jiehui,CAI Yuanlong.A picture quality evaluation method based on human perception[J].Acta Electronica Sinica,1999,27(4):79-82

[35] 金伟其,贾晓婷,高绍姝,等.彩色融合图像的质量主观评价[J].光学精密工程,2015,23(12):3465-3471
JIN Weiqi,JIA Xiaoting,GAO Shaoshu,et al.Subjective evaluation of quality for color fusion images[J].Optics and Precision Engineering,2015,23(12):3465-3471

[36] 陈锐,江奕辉.生成式 AI 的治理研究:以 ChatGPT 为例[J].科学学研究,2024,42(1):21-30
CHEN Rui,JIANG Yihui.A study of the governance of generative AI:taking ChatGPT as an example[J].Studies in Science of Science,2024,42(1):21-30

AIGC image quality evaluation indicators

XING Runmei[1,2] CHANG Shenglong[3,4] HE Kuan[5,6] ZHU Shuguang[5] GAO Qiong[5] HU Hao[5,7]

1 Logistics School,Henan College of Transportation,Zhengzhou 451460,China

2 School of Civil Engineering,Zhengzhou University,Zhengzhou 450001,China

3 College of Software,Henan Normal University,Xinxiang 453007,China

4 Henan Hengmao Chuangyuan Technology Co.,Ltd.,Zhengzhou 450016,China

5 College of Surveying and Mapping Engineering,Yellow River Conservancy Technical Institute,Kaifeng 475004,China

6 School of Surveying and Land Information Engineering,Henan Polytechnic University,Jiaozuo 454000,China

7 School of Water Conservancy,North China University of Water Resources and Electric Power,Zhengzhou 450046,China

Abstract Artificial Intelligence Generated Content (AIGC) technology offers a wide range of information genera-

tion services.However,the accurate assessment of AIGC quality is a critical issue that needs to be addressed.This study delves into the quality of images generated by large models and their evaluation metrics.First,it summarizes common methods for evaluating AIGC from a technical perspective,such as deep learning and computer vision approaches.The study introduces the metrics used in these evaluation methods,including accuracy,relevance,consistency,and interpretability,and examines their performance in evaluating diverse generated content.Then,to demonstrate the practical application of these evaluation metrics,this study conducts an evaluation experiment using images generated by ERNIE Bot as an example.Objective evaluation of the generated images is carried out through quantitative metrics like histograms and noise counts,while subjective evaluation focuses on the overall coordination and aesthetic appeal of the images.Finally,by comparing the results of objective and subjective evaluations,this study identifies highly reliable metrics for evaluating the quality of AIGC images,including color bias,noise count,and psychological expectations.This research provides a theoretical foundation for evaluating the AIGC quality and verifies the effectiveness and reliability of a combined approach using both objective and subjective metrics for AIGC product evaluation through experimental results.

Key words AI generated content (AIGC);deep learning;computer vision;image;quality assessment