

大模型视觉领域应用

在 AI 技术快速发展的今天,单一模态大模型或者基于大模型驱动的单一平台往往难以满足复杂场景下的多样化需求。多模态协同成为一种高效且灵活的解决方案,通过将基于大模型驱动的不同平台组合应用,或者直接使用多模态大模型,可以充分发挥优势,实现更强大的功能覆盖和更精准的任务执行。

3.1 Photoshop 智能化改造

Adobe Photoshop 作为图像处理领域的标杆软件,在大模型时代迎来了智能化升级的契机。通过引入大模型协同工作,Photoshop 实现了从工具型软件向智能创作平台转型,具体实现方案采用了三层架构。

(1) 基础层:集成 Stable Diffusion 等文生图模型,支持通过自然语言描述直接生成设计元素。要实现该功能,用户可以安装 Auto-Photoshop-StableDiffusion-Plugin 来实现。

(2) 增强层:对接 CLIP 等视觉理解模型,实现智能修图建议和自动化背景替换。

(3) 决策层:连接 GPT-4o 等推理模型,根据用户设计意图提供完整的创作方案建议。

例如,当用户需要制作电商海报时,系统可以:①通过对话理解产品卖点和风格需求;②自动生成多个版式原型;③智能推荐配色方案和字体组合;④实时监测设计规范符合度。显然,这种协同工作模式能够使设计效率得到显著提升的同时,还降低了专业软件的使用门槛。接下来,本节将介绍一些具体的实践案例。

3.1.1 AI 工具的选择与安装

本书选择 StartAI PS 插件作为辅助工具,帮助读者更高效地完成图像处理任务。StartAI PS 插件集成了多种智能功能,如一键抠图、智能调色和批量处理等,极大地简化了 Photoshop 的操作流程。无论是专业设计师还是初学者都能通过这款插件快速实现复杂的图像编辑效果。此外,插件还提供了详细的使用教程和案例演示,确保用户能够充分地利用其强大功能。接下来,本节将简单讲解 StartAI PS 插件的安装过程。

首先从 StartAI 官网下载对应操作系统的应用程序,目前官方提供了 Windows 版本和

macOS 版本；下载完成后双击应用程序进行安装即可，没有过多复杂的操作，不过，在安装过程中需要关闭 Photoshop，安装完成后可以通过如下几种方式启动 StartAI 插件：

- (1) 从桌面直接单击 StartAI 图标进行启动，启动 StartAI 时会自动启动 Photoshop。
- (2) 打开 Photoshop 时会默认启动 StartAI，此时在 Photoshop 桌面有 StartAI 图标。
- (3) 如果不小心将 StartAI 关闭了，也没关系，可以通过单击菜单栏增效工具中的 StartAI 调出该插件，StartAI 调出后如图 3-1 所示。



图 3-1 StartAI 安装好后调出的界面

注意 由于 StartAI 是一款收费软件，因此很多功能需要缴费才能使用，不过读者可以添加官方微信获取一定量的积分来测试相关功能。

3.1.2 面向行业的工具应用

在大模型时代，AI 的角色已从单纯的技术工具演变为能够为客户交付完整产品的解决方案。这种转变的核心在于 AI 产品需要深度融入行业场景，精准解决特定领域的核心问题，例如，在医疗健康领域，辅助诊断系统可直接提高影像识别准确率，治疗方案优化算法则能根据患者数据推荐最佳方案等。当然，这种从工具到产品的转变对 AI 开发者提出了更高要求：不仅需要扎实的技术能力，更要具备深刻的行业洞察力，确保开发的产品能切实创造商业价值并推动行业创新发展。接下来，本节将以 Photoshop 为基础工具，从行业的视角

来解读几个已具备客户交付雏形的工具应用。

1. AI 商品图

AI 商品图生成是给定一个商品原型图片，AI 算法会参考背景图像或者直接根据提示词生成新的商品展示场景。通过深度学习技术，AI 能够自动识别商品主体，并将其无缝融合到不同的背景中，或者根据用户需求创造出全新的虚拟场景。无论是电商平台的产品展示，还是广告设计的创意呈现都能够通过 AI 商品图生成来完成，大大节省了传统拍摄和后期处理的时间与成本。此外，AI 还能根据市场趋势和用户偏好，智能推荐最适合的展示风格，提升商品的吸引力和转化率。以图 3-2(a)所示香水为例，笔者并没有提供参考图像，而是选择输入了一段如下所示的提示词：背景是一间华丽屋子，商品放置在一张桌子上，但注意，背景不能遮盖商品。随后单击图 3-2(c)所示的“立即生成”按钮，在等待几分钟后，便生成了如图 3-2(b)所示的商品图。



图 3-2 AI 商品图生成

2. AI 换色

利用 AI 技术可以实现图像前背景的自动替换,通过深度学习算法分析图像内容,智能识别前景主体与背景区域。系统采用语义分割技术精确区分不同物体,再结合色彩迁移算法将背景色调与前景自然融合。用户只需上传图片并选择目标区域和希望替换的颜色,AI 便可在数秒内生成毫无违和感的合成效果图,同时保留原图的光影细节和纹理特征。笔者登记照如图 3-3(a)所示,此时的衬衣是白色的,换成蓝色衬衣的登记照如图 3-3(b)所示,设置界面如图 3-3(c)所示。



彩图



(c) AI换色操作界面

图 3-3 AI 换色前后对比

3. 产品精修

产品精修本质上就是对产品进行精细化处理,通过调整光影、色彩、材质等细节,提升整体视觉效果,使其更符合品牌调性和市场需求。以面部精修为例,通常会先进行基础瑕疵修复,如去除痘痘、细纹等,然后调整肤色均匀度,增强五官立体感。接着通过光影重塑突出产品质感,例如在护肤品精修中强化玻璃瓶身的高光折射,最后进行整体色调统一,确保在不同终端设备上呈现一致的视觉体验。专业精修师还会根据产品定位选择不同的风格,如高端化妆品采用冷调金属质感,母婴用品则偏向柔和的暖色调处理。基于大模型的 AI 产品精修则是将这些功能进行了融合,通过一键精修的方式获得了最终效果,同样以笔者登记照为例,原图如图 3-4(a)所示,精修后的图片如图 3-4(b)所示,操作界面如图 3-4(c)所示。



图 3-4 AI 精修前后对比

3.1.3 基于 AI 的图像生成

近年来,随着深度学习技术的快速发展,基于生成对抗网络(GAN)和扩散模型(Diffusion Model)的 AI 图像生成技术取得了突破性进展。这些技术能够根据文本描述自动生成高质量、高分辨率的图像,在艺术创作、广告设计、游戏开发等领域展现出巨大潜力,其中,Stable Diffusion、DALL-E 等先进模型通过大规模多模态数据训练,实现了对复杂语义的精准理解和视觉呈现。当前研究重点包括提升生成图像的细节质量、增强可控性及解决伦理版权等问题,未来有望推动人机协同创作的新范式。接下来,本节将分别从文生图和图生图两个角度来展示基于 AI 的图像生成。

1. 文生图

文生图技术通过输入自然语言描述直接生成对应图像,其核心在于文本编码器与图像解码器的协同优化。以 CLIP 为代表的跨模态预训练模型实现了文本与图像特征的精准对齐,使系统能够理解“戴着太空头盔的柯基犬在月球上冲浪”等复杂语义。最新研究通过引入动态阈值采样、注意力机制优化等技术,显著改善了生成图像的手部细节、文字渲染等传统难点。同样以 StartAI 为例,官方提供了大量文生图模型供用户选择,如图 3-5 所示,如果用户想要选择不同风格的模型,则可以先单击“更换模型”按钮,然后在右侧弹出的界面中选择。当然,如果是大模型领域的开发者,则可以寻找开源模型进行训练。



图 3-5 文生图模型选择和切换

选定模型后,用户还可以根据自己的喜好来选择风格。需要特别注意的是,这里提供了选择用户自己训练的 LoRA 模型的功能,如图 3-6 所示。



图 3-6 文生图风格选择和切换

完成模型选择和风格选择后,可以按照顺序编写提示词,此时 StartAI 的官方提供了一个较好用的功能,也就是提示词生成器和 DeepSeek 扩写功能,当然,用户如果不想要用这个功能,或者没有购买该插件,则可以通过大模型交互平台来生成提示词,类似地,在智谱清言和通义千问等平台都提供了生成提示词功能。本节以“一只小猫,趴在篮球场,晒着太阳”提示词为原点来生成,根据此场景应用提示词生成器,配置和生成过程如图 3-7 所示。

完成提示词的配置后,再进一步设置生成图像的尺寸,原始的图像尺寸为 512×512 像素,用户可以根据自己的需求进行调整,除此之外,用户还可以选择生成图片的数量和一些高级设置,本书对此不做过多陈述,用户可以自行探索。全部设置完成后,用户可以单击“立即生成”按钮,生成结果如图 3-8 所示。

2. 图生图

图生图技术作为计算机视觉领域的重要分支,主要通过语义解构和风格迁移实现图像到图像的转换。该技术在多个应用场景展现出显著价值,例如老照片修复、动漫化风格转换等视觉创作任务。接下来,本节以一幅简笔画为基础创作一幅 3D 风格绘画。原始简笔画如图 3-9(a)所示,新生成的 3D 风格图片如图 3-9(b)所示。

除了上述由一种形式的图像生成另一种形式的图像外,图生图也可以用于风格变换,例如将图 3-8 所示的照片风格转变为水墨风格,效果如图 3-10(b)所示。

大模型在图像处理领域展现出强大的多任务处理能力,主要包括 3 个层面的应用:基础图像处理、跨模态理解和生成式应用。通过微调预训练模型,可高效地完成图像增强、风



图 3-7 提示词生成器的配置界面



图 3-8 以“一只小猫,趴在篮球场,晒着太阳”生成的图像

格迁移、超分辨率重建等基础任务,特别适用于医学影像分析、卫星图像解译等专业领域,能够针对特定场景优化处理效果。多模态大模型实现了图文跨模态理解,支持智能修图、自动配文等创新应用,显著地提升了数字内容创作的效率和质量。此外,大模型还能实现基于AI的图像生成,为创意设计提供更多可能性。

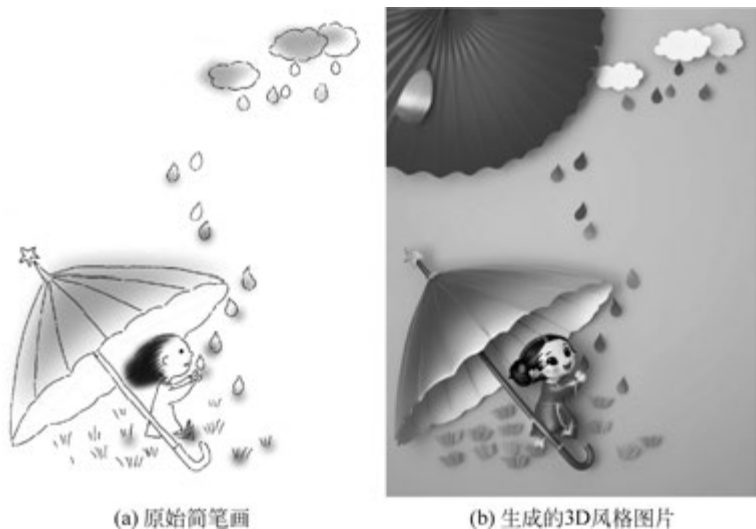


图 3-9 图生图之素描画生成 3D 风格画



图 3-10 图生图之风格变换

3.2 视频创作革命

基于大模型的视频创作正在颠覆传统影视制作流程,让每个人都能成为导演。通过自然语言输入,AI 可以自动地生成分镜脚本、匹配素材库中的画面,甚至实时渲染出高质量视频。这项技术正在短视频、广告和教育领域快速落地,大幅降低了专业视频制作的门槛。未来,结合多模态理解能力,AI 将实现从文字到 4K 视频的一键生成,彻底改变内容生成与传播方式。

3.2.1 一段话生成一段视频

一段话生成一段视频是指用户只需输入“落日余晖下的海边漫步”或“未来城市空中追

逐战”等文字描述,系统即可生成包含运镜、转场和配乐的完整视频,当然,部分功能可能由于模型和平台不同而存在差异。该功能通过语义解析自动拆分场景要素,调用 3D 模型库动态合成符合物理规律的光影与动作,并支持风格化参数调节。目前已有平台实现 1080P 商业级短片输出,其智能补帧技术可使 30s 视频的渲染时间从传统制作的 8h 压缩至 90s。本节将以智谱清言提供的清影-AI 生视频向读者快速展示如何实现一段话直接生成一段视频。图 3-11 是清影-AI 生视频的交互界面(<https://chatglm.cn/video?lang=zh>)。



图 3-11 清影-AI 生视频的交互界面

清影-AI 生视频是一款基于 AI 技术的视频生成工具,能够通过简单的文字描述或图片输入来快速生成高质量、富有创意的视频内容。它支持多种风格转换,包括电影级特效、动画渲染和 3D 建模,适用于广告制作、社交媒体内容创作及教育培训等多个领域。用户只需上传素材或输入创意思法,系统便可自动完成剪辑、配乐和特效合成,大幅提升视频制作效率。由于清影-AI 生视频目前只提供 5s 的视频生成,因此笔者就以“两个人在落日余晖下海边漫步”为输入,来生成一段视频,生成的结果如图 3-12 所示。



图 3-12 由清影-AI 生视频生成的视频

这里生成的视频可以下载、分享或收藏,当然如果用户觉得不满意,则可以让大模型重新生成。除了可以通过一段文本直接生成视频外,也可以通过输入的图像和文本来生成一段视频,例如可以先将如图 3-13 所示的图像上传,然后配以提示词“人跑向镜头”,看一看效果是怎样的,生成视频中的一帧如图 3-14 所示。



图 3-13 参考图像



图 3-14 基于参考图像和提示词由清影-AI 生视频生成的视频

除了智谱清言具备文生视频的能力外,Pika、Runway 和 Synthesia 等平台也提供了强大的 AI 视频生成功能。Pika 以其简洁的界面和快速的渲染速度受到用户青睐,Runway 则凭借多模态 AI 模型支持从文本到视频的一站式创作,而 Synthesia 专注于生成带数字人主

播的营销视频,能自动合成多语种配音与口型同步。此外,Stable Diffusion 的开发者推出的 SVD 模型进一步地降低了动态视频生成的技术门槛,使普通用户也能通过简单提示词生成流畅的动画片段。除了这些外,像 OpenAI 发布的 Sora 是文生视频的开山鼻祖,用户可以根据自己的需要和能够获得的资源来选择合适的产品和平台,而对于资深开发人员,也可以下载及部署一些开源的文生图大模型,以便进一步地测试效果。

3.2.2 DeepSeek 与剪映结合生成创意视频

除了可以通过上面基于文生图平台直接生成视频外,也可以通过 DeepSeek 与剪映相结合的方式打造创意视频。具体操作流程如下:首先利用 DeepSeek 强大的 AI 生成能力创作优质文案和分镜脚本,然后通过剪映的智能剪辑功能快速实现视频合成。这种组合方式既能保证内容创意质量,又能大幅度地提升视频制作效率。关键步骤如下:

(1) 用 DeepSeek 生成视频文案,如图 3-15 所示。



图 3-15 用 DeepSeek 生成视频文案

(2) 复制图 3-15 中的文本内容,粘贴到“剪映”→“AI 生成”→“视频生成”→“文本生视频”中,如图 3-16 所示。

(3) 选择生成模型,例如在图 3-16 的示例中,选择的是 Seedance v10,需要注意的是,不同的模型生成的效果不同,用户在使用过程中可以多尝试几次。

完成设置后,可以单击“开始生成”按钮,静待几分钟,便会生成视频,但此时生成的视频是没有声音的,用户可以使用其他专门的配音软件完成配音工作,并将配音与视频进行剪辑



图 3-16 将视频文案复制后粘贴到剪映中

合成。这种工作流可将传统以天为单位的制作周期压缩至几小时内完成,同时保证内容的新鲜度和平台适配性。最终生成的视频效果如图 3-17 所示。



图 3-17 剪映中生成的视频

笔者对生成的视频进行了深入分析,发现当前技术生成的视频在实际生产应用中仍存在较大局限性。虽然如 3.2.1 节所示,基于简单文本描述生成的视频可以满足某些低要求场景,但对于具有明确需求的特定应用场景而言,这项技术仍需经历长期的发展与完善过程。

3.3 3D 设计新业态

大模型技术正在重塑 3D 设计领域,其多模态生成能力覆盖文字、图片、视频及 3D 模型的全方位创作。这项技术显著地降低了专业设计门槛,具体体现在以下 3 个维度:

(1) 设计师通过输入文字描述或草图,即可获得高精度 3D 模型原型,支持实时编辑与多风格切换。

(2) 基于此技术衍生的“云建模”服务平台,通过自然语言交互即可完成复杂建模工作,正在重构游戏开发、影视特效等行业的创作流程。

(3) AI 生成的参数化模型能与 3D 打印、虚拟现实系统无缝对接,构建起从创意构思到实体产品的完整创新链条。

接下来,本节将从文生 3D、图生 3D 和模型生 3D 这 3 个维度向读者展示如何构建一个 3D 模型,在此过程中,作者使用了腾讯发布的混元 3D 大模型。当然,读者也可以选择其他的基座模型完成该任务。

3.3.1 一段文本生成一个 3D 模型

文生 3D 技术是指通过文本描述自动生成 3D 模型的创新过程。该技术主要包含 3 个核心环节:首先,系统对输入文本进行语义解析,理解其中的三维空间关系;其次,基于深度学习算法构建对应的几何结构;最后,自动生成材质贴图 and 场景布局。这项技术深度融合了 NLP、计算机图形学和深度学习三大领域,实现了从抽象文字到可视化立体模型的智能转化,在游戏开发、虚拟现实、工业设计等多个领域具有重要应用价值。腾讯混元 3D 的文生 3D 界面如图 3-18 所示。

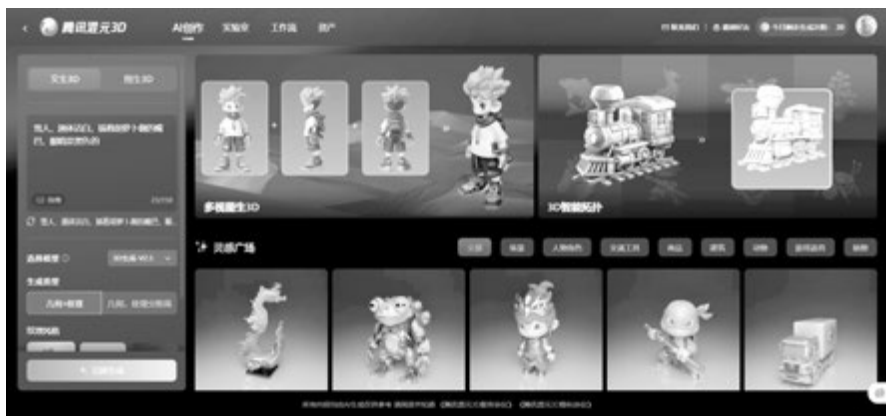


图 3-18 文生 3D 主界面

在图 3-18 的左侧输入框中,用户输入自己想要生成的 3D 造型,这里我们输入的是:雪人,通体洁白,插着胡萝卜做的鼻子,眼睛是黑色的,然后选择大模型和生成类型,以及纹理风格,最后单击“立即生成”按钮即可,生成后的效果如图 3-19 所示。



图 3-19 由一段文本生成的 3D 模型

图 3-19 展示了系统一次性生成的 4 个 3D 模型效果图。这些模型各具特色,用户可以根据个人偏好进行选择。经过笔者实际体验,第 1 个模型在视觉效果和功能设计上表现最突出。单击第 1 个模型对应的位置后,系统呈现的界面如图 3-20 所示。



图 3-20 3D 模型设置及下载界面

在图 3-20 中,用户可进行以下操作:首先,调整渲染灯光参数;其次,选择所需的保存格式;最后,单击“下载”按钮完成导出流程。

3.3.2 几张图片生成一个 3D 模型

除了可以通过一段文本来生成 3D 模型外,还可以通过上传一张或者几张(最好是 3 张以上)图片来生成一个 3D 模型。当然,这些图片最好能够 360°覆盖打算生成的物体,否则只能生成照片拍到的一侧。传统的方法是利用计算机视觉算法对这些图片进行分析和匹配,提取出物体的几何形状、纹理和深度信息。通过多视角立体视觉技术,先将这些 2D 图像转换为 3D 点云数据,再经过表面重建和优化处理,最终生成一个完整的 3D 模型,但到了大模型时代,这种情况发生了变化,用户只需上传照片,剩下的所有工作都可以交给大模型来完成。这种图生 3D 的方法特别适用于实物扫描和快速原型设计,能够高效地将现实世界的物体数字化。腾讯混元 3D 支持上传一张图片或者多张图片来生成 3D 模型,操作界面如图 3-21 所示。

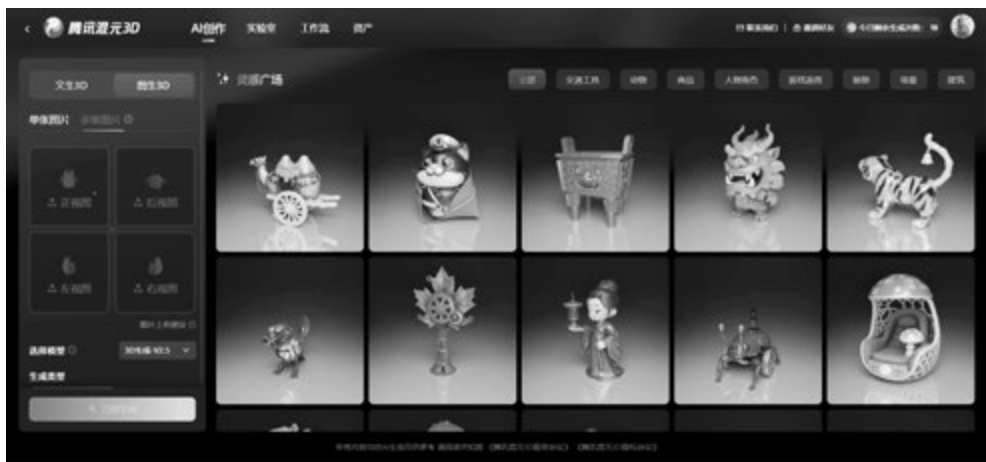


图 3-21 图生 3D 操作界面

本节将采用两种不同的图像输入方式生成 3D 模型:单张图像输入与多张图像输入。实验对象为笔者手边的保温杯,如图 3-22 所示,通过对比两种输入方式的建模效果,可以帮助读者直观地理解不同输入方式对 3D 建模质量的影响。

首先展示基于一张图片生成 3D 模型。将图 3-22 所示的图片上传至 3D 建模平台,在界面中选择合适的 3D 模型类型和生成参数。确认设置无误后,单击“立即生成”按钮,系统将自动将 2D 图片转换为 3D 模型,生成过程如图 3-23 所示,生成结果如图 3-24 所示。

最后相关设置和下载流程可参考图 3-20。从实验结果来看,基于单张图片生成的 3D 模型的效果显著优于传统计算机视觉方法,如图 3-22 所示,原始照片中水壶的 U 形把手并未完整拍摄,但图 3-24 所示的 3D 模型成功地还原了真实结构,这表明 3D 大模型已具备补全缺失信息的能力。值得注意的是,图 3-22 所示的背景信息并未被重构到最终模型中,这一特性在实际应用中具有重要意义。接下来,笔者利用 3 个角度拍摄的照片来重构图 3-22 中的水杯,从 3 个不同角度拍摄的照片如图 3-25 所示。



图 3-22 待进行 3D 生成的图片



图 3-23 单张图片生成 3D 模型的过程



图 3-24 单张图片最终生成的 3D 模型

从主观视角来看,笔者认为图 3-24 与图 3-26 的差异并不显著。这种相似性可能源于水壶结构的相对简单性。建议读者通过更多实例自行验证这一观察结果。

3.3.3 3D 模型生成对应拓扑模型

拓扑是三维建模中用于描述模型顶点与面编号结构的技术术语,其核心作用是通过固定顶点/面的编号顺序来确保修改器堆栈操作的可追溯性。在 3ds Max 等建模软件中,自动生成的拓扑结构会直接影响修改器对子对象的选择与记录方式。当模型顶点数量相同但



图 3-25 从 3 个不同角度拍摄的水壶照片



图 3-26 三张参考图片最终生成的 3D 模型

编号顺序不一致时,可能会导致变形过程中出现结构扭曲现象。拓扑技术通过显式子对象数据为相关修改器提供运算支持,例如“编辑网格”修改器就需要依赖拓扑信息才能正常运行。这项技术的实际应用价值在于能够重构低面数但高可用性的模型,有效解决高精度模型在动画制作过程中可能出现的兼容性问题。腾讯混元 3D 支持从文字、图像或者 3D 模型直接生成对应的拓扑图,但为了更加具有对比性,本节将使用 3.3.1 节和 3.3.2 节生成的 3D 模型作为输入,以此来生成对应的拓扑图。生成的结果分别如图 3-27 和图 3-28 所示。

3D 大模型具备强大的多模态生成能力,不仅能自动生成拓扑图,还能实现以下功能:一是生成 3D 动画、3D 纹理和 3D 游戏等数字内容;二是支持头像照片的 3D 角色生成,例如笔者上传登记照后,系统即可生成如图 3-29 所示的 3D 角色。

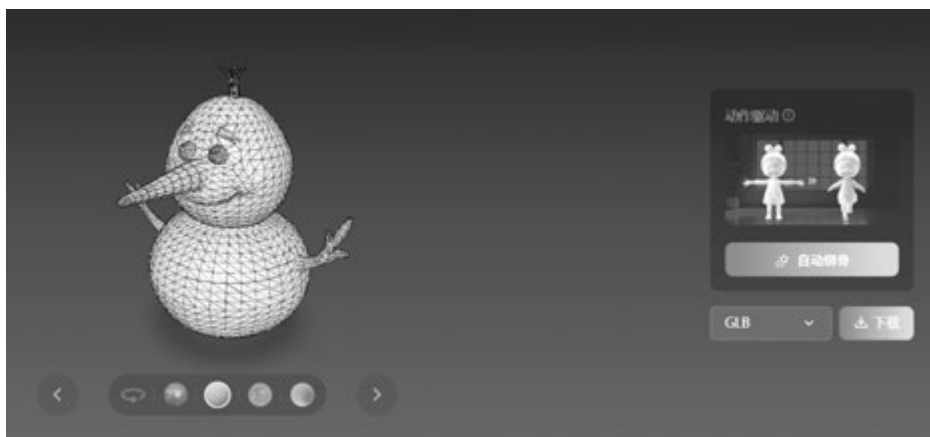


图 3-27 雪人的 3D 拓扑图



图 3-28 水壶的 3D 拓扑图

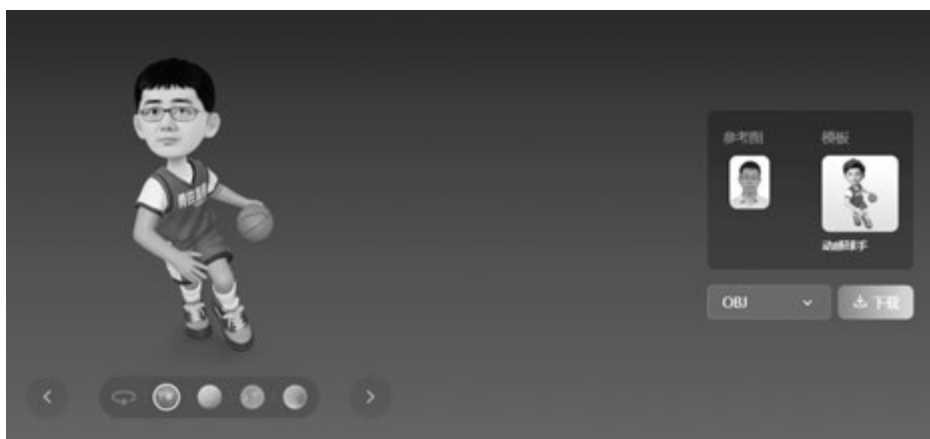


图 3-29 基于头像照片的 3D 角色生成

注意 需要说明的是,腾讯混元 3D(<https://3d.hunyuan.tencent.com/lab>)仅是本书选取的典型示例,但并非是唯一通道,实际上具备类似功能的平台和模型还有很多,读者可自行搜索以了解更多相关资源。

3.4 本章小结

本章围绕“多模态协同:大模型在视觉领域的应用”展开,详细探讨了在大模型时代如何通过多模态、多平台协作实现面向图像、视频和 3D 领域更高效、更灵活的任务执行,主要包括以下内容。

(1) Photoshop 智能化改造:通过引入大模型技术,Photoshop 从工具型软件转型为智能创作平台。具体展示了 AI 工具的选择与安装、面向行业的应用(如 AI 商品图生成、AI 换色、产品精修等),以及基于 AI 的图像生成技术(文生图、图生图)。

(2) 视频创作革命:基于大模型的视频创作技术正在颠覆传统影视制作流程。通过自然语言输入,AI 可以自动地生成分镜脚本、匹配素材库中的画面,甚至实时渲染高质量视频。案例展示了“一段话生成一段视频”的技术实现,以及 DeepSeek 与剪映结合生成创意视频的工作流程。

(3) 3D 设计新业态:大模型技术正在重塑 3D 设计领域,通过文生 3D、图生 3D 和模型生 3D 等技术,显著地降低了专业设计门槛。腾讯混元 3D 的案例展示了从文本或图片生成 3D 模型的全过程,并探讨了 3D 拓扑模型的应用价值。

本章的核心观点在于,多模态+多平台协作不仅能充分发挥各自的技术优势,还能为用户提供更高效、更灵活的解决方案。未来,随着大模型技术的进一步发展,多模态协同应用将成为推动行业创新的重要驱动力。