2152118 史君宝 期中报告

基于神经网络的生成式三维数字人研究综述: 表示、渲染与学习

1. 引言

近年来，在全球互联网行业“元宇宙”的概念都突然兴起，受到众多龙头企业的青睐。该概念起源于科幻小说《雪崩》，小说中人类可以通过虚拟化身进入

数字空间, 并相互交流，在当时引起了巨大轰动。

随着计算机视觉与图形学技术的发展，“元宇宙”已经走到互联网数字经济发展的最前沿，而虚拟数字人作为“元宇宙”技术的核心与基石，为了给予使用者沉浸式的体验，不仅要有逼真的外观，更要实现肢体动作和语言表达的流畅自然。未来随着人类使用者大规模涌入元宇宙，如何大规模生成高质量虚拟数字人成为重要问题。

数字人技术是指运用数字技术创造的，拥有数字化表现形式的虚拟人物。这些数字人往往以二维或者三维形式呈现，例如早期的平面动漫偶像，随着计算机图形学技术的发展，三维建模与渲染技术快速进步。一些顶尖的科研机构

和商业团队已经能够生产高仿真的虚拟数字人，比如近些年的一些虚拟偶像如初音未来等等。

但是，传统的三维数字人的建模过程包括形象采集、模型制作、纹理贴图、动作采集与驱动等复杂流程，依赖于专业的感知设备和精细的人力工作。不仅周期冗长而且只能定制化生产，无法满足大规模数字人的生产需求，阻碍了数字人的普及和应用。近年来，使用生成式人工智能技术创造数字人，以数据驱动的方式学习真实的数据分布，对数据采样生成新的样本表示，并对数据表示进行渲染从而呈现出高度真实的三维数字人，这种方法成为业内比较关注的新方法。

1. 生成式三维人建模流程

构建一个完整的生成式三维数字人模型主要包括 3 个步骤，分别为模型表示、渲染与学习。

首先，需要确定三维数字人模型的表示方式，常见的表示方式可以分为显式表示和隐式表示两种形式。显示表示是指直接给出满足条件的所有元素的集合，简单来说就是具体说明三维人模型每个线条的具体信息，尽管这项技术当前比较成熟，但是显式表示的精细程度会受到模型分辨率的限制，为了产生高拟真的数字人，需要大量元素来逼近模型的细节，从而造成模型复杂度的上升。而隐式表示仅需给出对于三维空间的某种约束，如符号距离函数、水平集等等。随着深度学习的发展，使用神经网络来逼近隐式函数，如深度符号距离函数、神经辐射场等。

之后是渲染，是从三维模型到二维图像的映射过程，决定了数字人呈现的视觉效果。由于我们难以模拟真实世界中复杂的光线，因此，传统图形学中的渲染过程一般是对物理世界成像原理的模拟和简化。这种简化虽然降低了计算复杂度，但是同时也造成了渲染质量的下降，使其难以生成高仿真的数字人形象。而采用神经网络渲染技术将数据驱动的神经网络与物理规律约束的渲染管线相结合, 能极大地提升了渲染的真实感。使生成高度拟真的三维数字人成为可能。

最后，生成式三维数字人模型需要对数据进行学习。由于三维扫描依赖专业的采集设备，成本高昂，因此难以构建大规模的数据集，采用二维图像数据采集更为方便。从有限小规模的扫描数据中学习到复杂多变的数字人模型成为一个重要难题，使用神经网络学习能在一定程度上解决上述问题。

1. 生成式三维人模型表示

3.1．基于显示表示的数字人模型

显式表示是三维物体的一类常见表示方法，即物体的三维结构被直接给出或通过参数映射的方式给出。三维人脸形变模型(3DMM)概念，是利用主成分分析从三维面部扫描数据得到一组表示面部形状和纹理的基，通过这组基的线性组合表示一个三维人脸。之后众多科学家不断丰富相关的数据集，以求更加丰富地表现三维人模型。

但是，考虑到线性3DMM表征能力有限，非线性 3DMM 的概念应运而生。区别于传统线性3DMM通过对形状和表情基底加权求和来表示三维人脸形状，一类非线性3DMM使用乘性模型或非线性变换来表示不同表情，达到了更强的表征能力。另一类方法借助可微渲染和深度神经网络，从二维人脸图像或视频中学习三维人脸先验，并使用训练好的解码器代替原始3DMM中的PCA基底，用于生成新的三维人脸。由于这类方法充分利用了大规模二维图片或视频进行学习，并结合了神经网络的非线性表示能力，因此其拥有比线性3DMM更强的表征能力，并且可以方便地应用于三维人脸重建等下游任务。

3.2．基于隐式表示的数字人模型

隐式表示使用一个连续函数来表示数据，因此隐式表示的复杂度不再取决于数据的空间分辨率，这为三维物体提供了一种灵活轻便的表示形式。然而物体是复杂多样的，通常难以用明确的函数对其进行准确的表征，因而研究者们常采用深度神经网络来逼近该复杂函数，即“隐式神经表示”。

在隐式神经表示中，隐式的连续函数通常采用多层感知机(MLP)进行参数化逼近，以空间三维坐标p=(x,y,z)∈R3作为输入，输出p点具有物理意义的属性fim(p)，形成了空间中的隐式场：F:p🡪fim(p)其中，fim(p)通常表示三维空间点的几何或纹理属性。

比如，基于生成对抗网络的三维数字人模型将隐式表示引入三维生成器，并通过判别器对数字人模型的二维渲染结果进行监督，从而生成真实自然的数字人。在学习过程中，从随机分布中采样一个隐空间编码(latent code)，同时结合三维坐标p=(x, y, z)和视角方向 d 作为输入，生成模型在三维空间中的特征属性，并进一步通过神经渲染器生成对应视角的二维图像, 作为判别器的输入. 判别器和生成器采用对抗损失函数迭代优化，不断优化生成器以假乱真的能力和判别器识别真假的能力，从而提升生成器的生成效果。

1. 生成式三维人渲染

4.1．传统图形学渲染

朗伯反射模型是计算机图形学中常用的一种对理想粗糙表面的漫反射模型，因其计算的便捷性被广泛应用于数字人的渲染中。在该模型中，观察者所接受的亮度不随观察视角变化，漫反射强度呈现各向同性的性质，遵循余弦定律，即漫反射强度取决于入射光线与其在物体表面的交点法向量的夹角，即I = I0cosθ，其中I是反射光强度，I0是入射光强度，θ是光线向量和表面法线向量的夹角。在渲染过程中，根据三维物体所处环境的光照条件计算出成像平面内该物体上各点对应的像素值。P(x)=A(x)ITS(x)，其中x表示坐标位置，P(x)表示图像像素颜色，A(x)表示图像本色，I表示图像像素该点的光线，S(x)表示图像像素点的法向量。

基于朗伯反射的漫反射分量在保持计算效率的情况下，可提供接近复杂模型的粗糙表面渲染效果。然而，在使用朗伯反射计算漫反射强度时未考虑物体微表面的粗糙程度。因此，虽然朗伯反射模型作为基础漫反射模型拥有较快的计算效率，但其最终渲染效果不够理想。

4.2．神经网络渲染

神经网络渲染是近年来新兴的一种结合深度神经网络的渲染方式，将给定的渲染参数输入网络进行推理，获取存储于神经网络中的隐式三维信息，再根据对应的图形学渲染原理进行合成，最终输出渲染结果。与传统图形学渲染中三维模型与渲染算法相互独立不同，神经渲染中的三维模型的信息存储于训练完成的神经网络之中，网络所学习的三维物体信息与其采用的对应渲染算法具有稳定的对应关系。同时，由于神经渲染所采用的底层渲染算法多基于光线物理传播过程建模，在复杂场景下神经渲染可达到逼真的渲染效果。

例如，神经体积渲染是一种将深度神经网络与计算机图形学中的光线投射算法相结合的渲染技术，将空间中各点的色彩与密度信息以隐式表示的形式存储于深度神经网络中。与显式的体积渲染中存储的离散体素集合不同，神经体渲染中存储的信息为三维空间中的连续场。

1. 生成式三维人学习

5.1. 显示模型学习

显式模型学习主要研究生成式数字人的参数化表示方法，以实现用低维的参数控制多变的数字人网格形状、纹理、表情，这类方法分为线性参数学习和非线性参数学习。

线性参数指模型的变化可以由一组基底的线性组合表示，为了获取低维线性控制变量，常使用PCA从三维扫描数据出发，利用特征值分解得到数据中线性无关的分量。假设 X = {x1,x2,……,xn}包含n个拓扑结构一致的三维网格，首先对X计算出平均值并归一化，之后对其协方差矩阵进行特征值分解，取特征值最大的前d个特征向量构成一组线性基底B，通过基向量的线性组合就可以生成新的数据。

5.2. 隐式模型学习

隐式学习方法一般从隐空间编码出发，优化神经网络空间中的三维位置到数字人隐式表示的映射。根据数据类型的不同，其学习方法也有所区别。、

基于三维数据的隐式学习方法通常通过自动解码器对隐空间编码进行优化来实现数字人的生成。在这类方法中，三维空间中的采样点p通过神经网络 fθ(·)表示为隐式结构，如符号距离场或者占据场，并通过计算与其三维真值 gt(p)之间的距离来训练，主要学习目标可以表示为L=∑||gt(p)−f(p)||。

基于二维数据的隐式学习方法通常从隐空间采样出发学习数字人的三维隐式表示，并利用隐式渲染方法将三维数字人投射到二维空间，再通过多种损失函数来比较真实数据与生成数据之间的差异来实现学习。主流工作采用生成对抗的学习方式，从隐空间中进行采样，并利用生成器G和判别器D的对抗学习实现模型的训练。具体而言，生成器从随机变量z生成三维数字人，同时训练判别器D来区分真实数据Igt与生成结果G(z)。此外，常在GAN中加入R1正则化损失函数以实现模型的稳定训练，优化目标可以表示为L=E[f(−D(Igt))− λ||∇D(Igt)||2]+E[f(D(G(z)))]，其中f(x)=−log(1+exp(−x))，λ为 R1正则化的系数。

7. 总结与反思

目前绝大部分生成式数字人模型都是基于二维图像或者三维表面数据进行学习的，通过对人体表面的观测结合简化的先验约束来推断数字人的变化与运动规律。但是现有的数字人模型仍然和真人存在较大差距，相关数据还需要医学工作者的研究和漫长的采集过程。

总的来说，生成式数字人的表示形式正从显式表示向隐式表示的方向发展，渲染方式正从传统渲染方法过渡到神经网络渲染，学习方法越来越倾向自监督、对抗学习等弱标注场景。现有数字人技术离实现“元宇宙”中高度真实、自然交互的虚拟形象仍存在巨大差距，期待未来的继续研究。