分类号：TP391.4　　单位代码：10335
密　级：公开　　学　　号：

浙江大学

博士学位论文

中文论文题目：动态三维人体的隐式神经表示方法研究

英文论文题目：Implicit Neural Representations for Dynamic Human Bodies

申请人姓名：彭思达
指导教师：周晓巍
学科 (专业)：计算机科学与技术
研究方向：三维计算机视觉
所在学院：计算机科学与技术

论文递交日期　二〇二三年四月十二日

致谢

二十一年的学生生涯结束了。我很幸运，能一路上持续地享受学习的快乐，并最终写下这一份博士学位论文，作为个人学业成果的总结。时常庆幸自己能较为纯粹地热爱学习，在很大程度上单纯地享受着获取知识和自我提升带来的快乐。很感谢各个阶段的老师、我的父母、我的妻子、一路上认识的同学对我这方面性格的影响。

在高二之前，我对学习的意义其实没有什么概念，只是因为比较乖、听家长老师的话，按部就班地走着。老师布置作业，我就认真完成，如果要考试了，我就好好复习。在做完被布置的事情以后，就找点其他事情打发时间。如果一直是这种状态，我很可能在大学的时候会荒废自己的时间，开始蹉跎岁月了。很庆幸在高二的时候遇到冯淑珍老师这么好的班主任，她让我意识到学习的意义，让我对更好的自己有了期盼，也让我在高二、高三两年里养成了持续学习的习惯。冯老师经常找我谈话，聊日常学习、聊人生规划，无偿地额外花时间帮我查缺补漏。我永远忘不了她在每次和我谈话以后对我期盼的目光，感觉自己真的被寄予厚望。作为教育的受益者，冯老师对我的培养以及我后续的成长让我深深地信仰教书育人的价值。有人说：“教育是一个灵魂唤醒另一个灵魂。”在我身上，是冯老师唤醒了我，以她无条件的好培养了我的学习观。人生中很多事情可能是没有意义的，但我非常确信，学习和教育肯定是有意义的。

到了本科，课余时间变得很多，一周经常有好几天以上可以自由分配的时间，我拿这些时间学了很多课外的知识。记得第一次接触编程的时候，发现用几行代码就能在电脑上显示酷炫的效果，让我很震撼。因此我本科一直有个愿望是搞清楚操作系统的运行原理，后面也确实学了很多这方面的知识，为我科研的实验能力打下了基础。本科一个非常正确的决定是加入工高班，这是一个技术极客扎堆的地方，我遇到了很多趣味相投的人，比如利军、栗橙、晓刚。记得大二的时候经常和利军、栗橙一起去自习，然后晚上十点多一起回宿舍。我很怀念那个时候。那段时间学课外知识真的完全凭兴趣，甚至没有什么目标，只是觉得什么有意思就学什么，没有压力。虽然本科前三年学了挺多东西，但比较可惜的是缺少一个好的向导，导致没有学得很深入的方向。

很幸运的是，大四上我遇到了求学路上另一位重要的老师，周晓巍老师。在接下来的几年里，我在他的引导下进行了深入的系统性科研训练。教育在我这一段经历中再一次闪耀起了光辉，冯淑珍老师培养了我持续学习的习惯，而周晓巍老师培养了我解决技

术难题的科研能力和性格品质。周老师不仅是我的科研导师，更是我的人生导师。坚持是他教给我的一个珍贵品质。记得刚认识的那一段时间，他常和我谈坚持的重要性，他让我不用担心现在没啥基础，只要坚持做科研，总能变得很厉害。我通过博士这几年的实践深刻理解了坚持的含义。学习过程中经常会遇到很难的算法，我一开始会畏难，后来我发现，只要坚持去学习这个算法，学个几天或者一个月，原本很难的算法真的慢慢变得简单、变得形象、变得亲切。我读博期间反复经历了这样的体验，让我惊讶于人脑的神奇和坚持的可贵。做实验、写论文也是这样。开始一个新项目时的未知感或工作量可能让人畏惧，但只要我规划有度，每天坚持去完成一部分内容，就能体会到“进一步有进一步的欢喜”。在实验室管理方面，我常常感概周老师是个有大智慧的人，想着自己何时才能达到他的境界。他会制定一个明确的科研目标，稳步推动着实验室的进展，不会急于一时地去完成。有时候我们进度慢了或者没做，他从来不会说我们什么，而是再次规划要做的事情和时间点，似乎一切都在把握之中，最后也确实大多完成了。周老师在学生面前一直情绪稳定，总是有条有理、不急躁地做科研。我感觉能做到这一点真的很难，特别是对于一个青年教师而言。周老师会是我一直学习的目标。

感谢鲍虎军老师建立了这样世界一流的科研团队，让我有机会能在团队中学习科研，接受顶级的科研训练。感谢刘缘在我第一个科研项目中的帮助，他让我见识到了怎样是靠谱的合作者。在博士期间我们进行了多次合作，刘缘总是能让我不由自主地佩服他的能力之强和效率之高。感谢实验室的学弟学妹们，在和他们的交流讨论中我完善了自己的科研方法，感受到了与顶尖学生相处的快乐。

感谢我的父母，他们总是很信任我，让我有勇气去完成这二十一年的求学之路。感谢我的妻子，她教会了我如何更好地与人相处、如何去关心人和爱人。

摘要

从观测视频中重建动态三维人体表示是计算机视觉和图形学领域的前沿热点问题，是数字内容制作、远程虚拟会议、影视制作等应用的重要技术。传统的人体建模技术可以构建高精度的数字人体模型，但这些方法依赖于复杂的硬件设备，比如深度相机、稠密相机阵列，限制了这些工作的使用场景，并且提高了建模成本和用户门槛。近年来，神经辐射场展现了从观测图片中重建高质量三维场景的能力。但此类方法需要稠密视角图片的输入，并且无法建模可驱动的动态人体模型。除此之外，此类方法的渲染速度较慢，无法满足实时应用的需求。

基于多视图几何理论与深度学习方法相融合的思想，本文提出了一系列面向动态人体建模与渲染的隐式神经表示方法，致力于解决稀疏视角建模、可驱动人体模型、几何表面重建、实时渲染这四个人体建模领域的关键问题，实现了从稀疏视角视频中创建具有高质量的可驱动人体模型。本文主要的研究成果如下：

（1）针对从稀疏视角视频重建动态三维人体模型的问题，本文提出了一种基于结构化隐变量的人体神经辐射场表示，可以有效地整合输入视频中不同时刻的观测信息。实验结果表明本方法可以从稀疏视角视频甚至单目视频中重建高质量的三维人体。

（2）针对可驱动的数字人建模问题，本文提出了一种基于骨骼蒙皮驱动的人体神经辐射场表示，将动态人体建模为空间变形场和标准空间下的神经辐射场。本文在 Human3.6M 和 ZJU-MoCap 数据集上验证了该方法的有效性。

（3）针对从视频中重建高质量人体几何的问题，本文提出了一种基于符号距离场的动态人体几何表示，利用程函方程对几何优化过程施加正则化。在多个数据集上的实验结果表明，本方法在人体几何重建方面大幅度地超过了之前的方法。

（4）针对动态人体的实时渲染问题，本文提出了一种基于多层感知机图的动态场景表示，通过一组小型多层感知机网络建模三维场景，从而降低了网络的推理成本，提升了渲染速度。在 NHR 和 ZJU-MoCap 数据集上的实验结果表明，本方法在渲染速度方面远远超过了之前的方法，并且在渲染质量上表现出了最好的效果。

关键词：三维人体建模，隐式神经表示，神经渲染

Abstract

Reconstructing dynamic human body representations from RGB videos is an important problem in computer vision and graphics, which has many applications, such as digital content creation, immersive telepresence, and movie production. Traditional methods have shown impressive results in reconstructing human geometry and appearance, but they rely on complex hardware, such as depth sensors and dense camera arrays, which limits their applications in real-world scenarios and improves the modeling cost. Recently, neural radiance fields have exhibited great performance on reconstructing 3D scenes from observed images. However, such methods require dense input views and cannot handle dynamic humans. Moreover, its rendering speed is slow, which is not suitable for real-time applications.

Based on the theory of computer vision, graphics and deep learning, we proposes several implicit neural representations for dynamic human bodies, which aim to solve the key problems of sparse-view modeling, animatable human models, geometry surface reconstruction, and real-time rendering. With the proposed representations, we are able to create high-quality animatable human models from sparse-view videos. The main contributions of this paper are as follows:

(1) To address the problem of reconstructing dynamic 3D human models from sparse-view videos, we propose a novel neural human neural field based on structured latent variables, which effectively integrates the observed information across video frames. Experiments show that our method can reconstruct high-quality 3D human models from sparse-view videos.

(2) To address the challenge of modeling animatable digital humans from videos, we propose a novel neural human radiance field based on the skeleton-driven deformation framework, which models a dynamic human as a neural radiance field in the canonical space and a deformation field. Experiments on the Human3.6M and ZJU-MoCap datasets show that our method not only achieves high-quality novel view synthesis, but also outperforms previous methods by a large margin in terms of novel pose synthesis.

(3) To reconstruct the high-quality geometry of human bodies, we propose a novel dynamic geometry representation based on signed distance fields (SDF), which regularizes the optimization process with the Eikonal loss. Experiments on multiple datasets demonstrate that

our method significantly outperforms previous methods in terms of geometry reconstruction.

(4) To achieve real-time rendering of dynamic human bodies, we propose a novel dynamic scene representation based on multilayer perceptron (MLP) maps, which models 3D scenes with a set of small MLP networks, thus reducing the inference cost of MLP networks and improving rendering speed. To validate our method, we conduct extensive experiments on the NHR and ZJU-MoCap datasets. Experimental results show that our method exhibits state-of-the-art performance in terms of rendering quality and speed.

Keywords：3D Human Modeling, Implicit Neural Representations, Neural Rendering

目录

图目录

表目录

第 1 章 绪论

1.1 研究的背景与意义

三维人体数字化技术致力于创建人们在数字空间中的虚拟化身。该虚拟化身不仅要有高质量的几何与外观，还要实现逼真自然的肢体表达。人体数字化是计算机视觉和图形学领域的前沿热点问题，具有广泛的应用，比如远程虚拟会议、影视制作、游戏建模、元宇宙、数字孪生、虚拟助手、数字内容制作等。图 1-1 展示了一些典型应用。Google Relightables[1] 利用稠密视角阵列重建人体几何与材质，从而能将数字人体放置于不同光照的数字场景中，为影视制作提供技术支持；Google Starline[2] 基于多视角深度相机实时地重建人体，将数字人体传输到远程，使得参会者能更立体地呈现在对方屏幕，提升了远程会议的体验；瞩深数字科技[3] 通过神经网络记录非物质文化遗产，实现了非物质文化遗产的数字化，为非物质文化遗产的保护提供了新的方式。三维人体的数字化技术也是我国在信息领域优先发展的技术之一。我国 2021 年印发的《十四五规划和 2035 远景目标纲要》将动态环境建模和实时动作捕捉列为发展数字经济的重点方向。

长久以来，研究人员一直在追求以更便捷的设备和更低的成本构建更高质量的虚拟数字人。传统的人体建模技术主要有两类，分别是基于三维重建的方法和基于图像插值的方法。之前的研究工作[4-7] 借助多个 RGB-D 相机拍摄目标人体的 RGB 图片和深度图片，然后使用多视角深度融合、表面重建和纹理贴图等技术获得带纹理的三维网格模型。虽然这些方法能实时地生成可渲染的网格模型，但多视角深度相机的使用限制了这些工作的应用场景，并且提高了使用难度和用户门槛。除此之外，由于贴图质量和几何精度有限，基于带纹理的三维网格模型进行渲染得到的图片往往不够真实。为了实现照

(a) 影视制作 (b) 全息视频会议 (c) 非物质文化遗产数字化

图 1-1 动态三维人体建模与渲染的相关应用。图片分别来自 Google Relightables[1]、Google Starline[2]、瞩深数字科技[3]。

片级的渲染，一些研究工作[1,8-9] 搭建了稠密相机阵列用于拍摄目标人体的多视角图片，并使用光场技术[10-11] 插值图片得到新视角下的图片。尽管这些方法能实现非常逼真的渲染，但这依赖于稠密的多视角图片，从而导致了巨大的存储和传输成本。

近年来，一些工作[12-15] 提出将三维场景建模为隐式神经表示，利用神经网络编码目标场景。隐式神经辐射场 NeRF[15]是其中的代表性工作。具体而言，该工作使用神经网络预测场景中任意三维点的体素密度和颜色，可以建模高分辨率的三维场景。尽管隐式神经辐射场[15]在三维场景建模上取得了很好的效果，这个技术在动态人体建模与渲染上仍然存在一些局限性。首先，神经辐射场只用神经网络编码了场景三维点在特定时刻的状态，导致其只能建模静态场景。除此之外，隐式神经辐射场的训练需要稠密的观测图片作为监督信号，才能从图片中恢复出高真实感的三维场景。然而，动态人体需要搭建稠密相机阵列才能获得稠密视角的观测图片。其次，用户通常有操作数字人的需求，用于表达一些肢体语言或者操作虚拟世界中的一些物体。目前隐式神经辐射场用一个神经网络表示三维场景，缺少显式的可操作性，因而无法表示可驱动的数字人。再次，隐式神经辐射场虽然能实现照片级的渲染，但重建出的三维场景几何一般比较粗糙，和真实几何有较大的差距。而一些数字人应用需要比较好的几何质量，比如重光照、游戏角色制作、虚拟换衣等应用。最后，渲染隐式神经辐射场需要大量次数的神经网络推理，以致于渲染过程比较慢。对于远程虚拟会议等应用，实时渲染是必备的功能。

针对现有动态三维人体建模和渲染所面临的问题与挑战，本文基于多视图几何理论与深度学习方法相融合的思想，提出了一系列人体数字化技术，实现了从稀疏视角视频中创建具有高质量的可驱动数字人体模型。具体而言，首先，为了从稀疏视角视频中建模数字人体，本文提出了基于结构化隐变量的神经辐射场，用于表示动态人体，并实现了时序信息的整合。其次，针对数字人体模型的可驱动性问题，本文研究了神经辐射场与骨骼蒙皮驱动模型的结合，实现从图片中优化得到蒙皮权重场，从而支持了通过人体姿态显式地操作基于神经辐射场的数字人体。然后，为了提升人体三维几何的重建质量，本文提出了基于符号距离场的动态人体几何表示，在网络训练过程中为几何优化提供了有效的约束，在保持高质量渲染的基础上，获得了高质量的人体几何模型。最后，针对动态场景的渲染问题，本文研究了神经辐射场的空间解耦，通过一组搭载小型神经网络的动态平面表示动态场景，从而降低了神经辐射场的网络推理成本，并以此为基础构建了一个面向动态场景的实时渲染流程。

1.2 研究目标与面临的挑战

本文致力于研究动态三维人体的建模与渲染技术，通过设计新颖的隐式神经表示，实现基于较为便捷的采集设备重建高精度的可驱动人体模型以及高效的动态场景渲染，从而促进虚拟数字人在元宇宙、远程虚拟会议、自由视角视频和虚拟助手等应用的落地。为了实现该目标，本文主要面临以下四个挑战：

（1）如何从稀疏视角视频重建动态三维人体：相比于稠密相机阵列，稀疏相机阵列易于部署且成本低，用户门槛较低，因此基于稀疏相机阵列建模数字人具有很高的实际价值。然而，即使是最近提出的隐式神经辐射场 NeRF，在该问题上也面临诸多挑战。首先是神经辐射场局限于静态场景的建模，无法表征动态人体。因为神经辐射场只预测场景三维点在特定时刻下的几何和外观，没有将时间作为输入变量，所以无法表示时变的场景。除此之外，稠密观测图片的输入是最优化神经辐射场的必要条件，而稀疏相机阵列只有少数相机视角，无法满足输入数据的条件，因此造成了另一个挑战。之所以优化过程依赖于稠密的观测视角，是因为三维场景投影到二维图片时丢失了深度信息，导致从二维图片恢复三维场景的过程存在歧义性，一张二维图片往往对应了多种可能的三维场景。稠密的多视角图片可以很大程度地消除歧义性。因此，如何设计动态人体表示以及如何消除稀疏视角下优化的歧义性是基于稀疏相机阵列建模数字人的两个关键难点。

（2）如何构建可驱动的人体模型：数字人模型的可驱动性是诸多数字人应用的关键，例如数字内容创作通常需要操作数字人完成特定的动作，或者社交应用中人们需要驱动数字化身表达肢体语言、与对方进行互动。传统的可驱动数字人大多是基于骨骼蒙皮驱动算法的三维网格模型。这样的人体模型主要有三个局限：首先，获取高精度的三维网格模型需要复杂硬件设备的支持；其次，带纹理的三维网格模型的渲染真实感有限；最后，建模师需要精心设计目标数字人的蒙皮权重以支持骨骼蒙皮驱动算法。考虑到神经辐射场易于从图片中建模、渲染真实感高等优点，设计基于神经辐射场的可驱动人体是可靠的发展方向。然而，神经辐射场使用单个全连接神经网络表示三维场景，而实现神经网络的可操作性并不直观。为此，本文将动态数字人解耦为标准姿态下的静态人体模型和任意姿态下的连续空间变形场，静态人体模型由神经辐射场建模，连续空间变形场由骨骼蒙皮驱动算法建模。该表示解决了神经辐射场可驱动性的问题。但是，基于骨骼蒙皮驱动算法表示连续的空间变形场带来了另一个挑战：为了驱动某一空间三维点，骨

骼蒙皮驱动算法需要已知该三维点的蒙皮权重，因此实现连续的空间变形场需要获得任意空间三维点的蒙皮权重。如何定义任意三维点的蒙皮权重和从图片中预测蒙皮权重是数字人自动化建模亟需解决的难点。

（3）如何从视频中重建高质量的人体几何：高精度的数字人几何模型是一些数字人应用的基础。例如影视制作大多会将虚拟数字人放置于新的环境，并根据环境光照进行数字人的重光照，实现数字人与新环境的融合。该应用要求数字人在拥有高真实感外观的同时也要有精细的三维几何，从而能支持传统图形学的渲染管线达成高质量的重光照。因此，从多视角视频中建模具有高精度几何的数字人在实际中有着很大的应用潜力。虽然神经辐射场可以从图片中恢复出支持高真实感渲染的三维场景模型，但该工作难以基于图片解耦几何与外观，导致从神经辐射场抽取得到的场景几何往往比较粗糙，而且空间中容易存在漂浮物。原因在于神经辐射场在优化过程中缺少对场景几何的约束，而且一些几何上的瑕疵容易被外观所弥补。如何有效约束由神经网络编码的三维几何是重建精细几何模型的关键难点。一个解决思路是引入符号距离场（Signed distance function）[16-18]，代替神经辐射场中的体素密度表示场景几何，并施加程函方程（Eikonal equation）来正则化符号距离场的优化过程。然而，如何从视频中学习基于符号距离场的动态数字人也是一个需要探索的研究内容。

（4）如何实现动态人体的实时渲染：面向动态场景的实时高真实感渲染，是实现沉浸式虚拟会议、可交互数字人、自由视角视频等应用的关键技术。然而，神经辐射场 NeRF 的渲染速度远远未能达到实时。其原因在于神经辐射场采用了八层全连接神经网络编码三维场景，保证神经网络有足够的容量记录场景的几何和外观。因为神经辐射场的体积渲染过程需要大量次数的网络推理，所以即使是八层全连接网络也会造成较慢的渲染速度。为了突破神经辐射场在渲染速度上的局限性，本文提出使用一组小型全连接网络共同编码动态场景，通过让每个小型全连接网络负责场景的每一小块区域，实现了高效且高质量的渲染。虽然该方案具有理论上的可行性，但在实际实现中需要存储大量的小型神经网络，导致存储成本高以及训练时间长。一种解决思路是使用一个超网络（Hypernetwork）[13,19] 记录大量的小型神经网络，但仍存在两个挑战：首先，简单地借助超网络预测大量小型神经网络将带来较大的计算成本，降低模型推理速度；其次，网络参数通常有非常高的维度，准确地编码网络参数要求超网络具有高性能的网络结构。因此，如何通过超网络高效且准确地预测小型神经网络是亟需研究的关键技术。

1.3 本文内容与结构

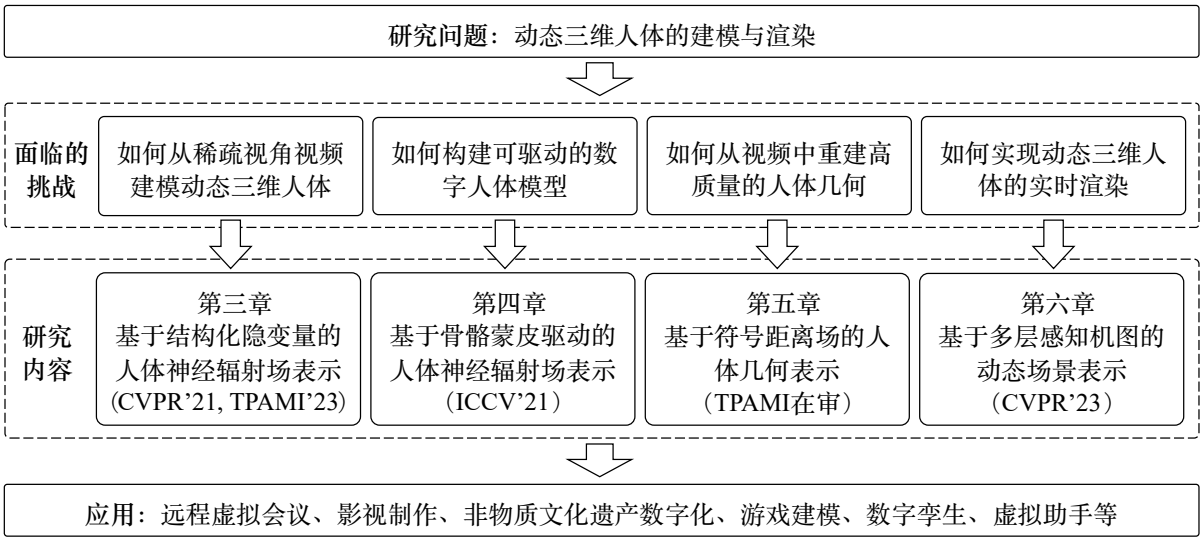


图 1-2 本文的内容与结构。

本文围绕动态人体建模与渲染展开研究，通过探索稀疏视角重建、可驱动数字人、几何建模、实时渲染这四个数字人领域的关键问题，实现从稀疏视角视频中重建高质量可驱动的动态人体模型，并支持动态场景的实时渲染。图 1-2 总结了本文主要的研究内容，分为：如何构建基于隐式神经表示的动态人体，并在训练中累积时序观测信息，实现从视频中重建支持高真实感渲染的人体模型；如何结合神经辐射场与骨骼蒙皮驱动算法，并有效稳定地从输入数据中自动获取基于隐式神经表示的蒙皮权重场，从而构建可驱动人体模型；如何基于符号距离场构建动态人体几何，并使用可微分体积渲染优化符号距离场，从而重建高精度的几何模型；如何高效地表示一组小型全连接神经网络，并通过超网络来准确地预测网络参数，从而以一组小型神经网络表示动态场景，降低网络预测的推理成本，实现实时且高真实感的渲染。本文在结构上分为七个章节，在本章介绍了研究的背景和意义和研究目标与面临的挑战。本文后续章节的主要内容如下：

第二章对相关的研究工作进行了综述。本章首先介绍了三维人体建模的传统技术，包括基于深度相机的方法和基于稠密视角阵列的方法。然后，本章介绍了基于深度学习的人体建模方法。此类方法通过在真实三维人体数据集上训练深度神经网络学习人体数据的先验，实现了从稀疏的人体观测数据中预测完整人体的三维几何和外观。最后，本章介绍了可微分渲染器，以及基于可微分渲染的静态场景和动态人体建模与渲染技术。

第三章提出了一种支持稀疏视角重建的动态人体隐式神经表示 Neural Body。该人

体表示基于隐变量模型（Latent variable model）假设不同视频帧中观察得到的人体模型可以被同一组隐变量编码，并通过可微分渲染优化这组隐变量以拟合不同视频帧中的人体图片，从而整合了视频的时序信息。具体而言，本章预定义了一组可学习的结构化隐变量，即这组隐变量固定在可变形人体参数模型的顶点上，其空间位置与人体参数模型的顶点位置绑定，随人体姿态的变化而变化。为了生成特定视频帧下的数字人体模型，本章首先从输入的稀疏视角图片中估计目标人体姿态，然后基于人体姿态对结构化隐变量的空间位置进行变换。最后，本章设计了一个基于三维卷积的神经网络，对变换位置后的结构化隐变量进行卷积编码，提取得到一个三维特征向量体，随后预测任意三维点的体素密度和颜色，即得到了目标视频帧下的数字人体模型。为了优化模型参数，本章通过可微分的体积渲染将数字人体模型投影为二维图片，并与目标视频帧下相应的观测图片做比对，通过最小化渲染误差以更新结构化隐变量与神经网络的参数。为了评估本章方法的有效性，本文创建了一个多视角数据集，称为 ZJU-MoCap。该数据集采集了多段动态人体进行复杂运动的视频。在所有采集的视频中，本章提出的方法在自由视角视频合成方面表现出最先进的性能。本章还在公开数据集 People-Snapshot[20] 上展示了上述方法可从单目 RGB 视频中重建支持高真实感渲染的数字人体模型。

第四章提出了一种基于骨骼蒙皮驱动算法的可驱动隐式神经辐射场 Animatable NeRF，用于表示可驱动数字人。本章将可驱动数字人分解为两个部分：标准坐标系下的人体模型和变换到任意人体姿态的空间变形场。本文使用隐式神经辐射场表示标准人体模型，而通过骨骼蒙皮驱动算法产生连续的变形场。具体而言，给定多视角视频，本方法从多视角图片中预测三维人体骨架，并使用一个多层全连接神经网络表示蒙皮权重场，用于预测任意一个三维点的蒙皮权重。对于世界坐标系下的任意一个三维点，本方法基于骨骼蒙皮驱动算法[21] 将蒙皮权重和三维人体骨架相结合，得到该三维点的变换矩阵，然后将其从世界坐标系变换到标准坐标系，最后索引标准坐标系下的隐式神经辐射场，得到数字人体模型的外观和几何。这种表示方法有两个优点。首先，由于三维人体骨骼易于预测[22]，因此不需要联合优化，从而减少了优化空间，并对变形场的学习提供有效的正则化。其次，通过在标准坐标系学习额外的蒙皮权重场，本方法实现通过输入人体骨架以显式地驱动神经辐射场。本章在 Human3.6M 数据集[23] 和第 3 章提出的数据集 ZJU-MoCap 上评估了上述所提出的方法。在数据集的所有视频序列上，本文的方法在新视图合成和新姿态合成方面表现出最先进的性能。

第五章提出了一种基于符号距离场的动态人体几何模型 Animatable SDF 来实现高质量几何重建。相比于第四章中在标准坐标系下用神经辐射场建模人体的方法，本章提出的模型使用符号距离场表示标准空间中的人体几何。与体素密度场相比，符号距离场在零水平集（Zero level set）处具有明确定义的表面，这有助于在优化人体几何的过程中施加直接的正则化。这里的一个挑战是如何从视频中学习标准标准坐标系下的符号距离场。球体追踪（Sphere tracing）[16,24] 是渲染符号距离场的经典方法。然而，由于世界坐标系和标准坐标系之间可能存在复杂的人体运动，因此很难在世界坐标系中沿着相机射线找到表面点。为了解决这个问题，本章使用骨骼蒙皮驱动算法计算标准坐标系和空间坐标系之间的非刚性变换，然后使用基于符号距离场的体积渲染技术[17] 将动态人体模型渲染到观测到的图片空间。本章提出的方法在单目视频和多视角视频中进行了评估，实验结果表明这种新颖的动态人体模型有效提升了几何重建的精度。

第六章提出了一种基于多层感知机（MLP）图的神经体积视频表示 Dynamic MLP Maps，实现了动态场景的实时渲染。本章提出的方法的关键思想是将动态场景的每一帧表示一组小型全连接神经网络，其参数存储在称为 MLP 图的二维图片中。这个 MLP 图由所有帧共享的二维卷积神经网络进行预测。具体而言，给定一个多视图视频，本渲染方法从中选择一组视图并将其输入到二维卷积编码器中，以获得每一个视频帧的隐变量，然后二维卷积解码器从隐变量中回归得到一张二维图片。该图片的每一个像素存储一个小型全连接神经网络的参数向量。本章将这样的二维图片称为 MLP 图。为了使用 MLP 图建模每一帧的三维场景，本方法将三维空间中的任意三维点投影到 MLP 图上，得到三维点在 MLP 图上的二维投影坐标，从而从 MLP 图索引得到相应的全连接神经网络参数，随后载入对应的网络结构，最后预测三维点的体素密度和颜色。本渲染方法使用一组小型全连接神经网络表示三维场景降低了网络推理成本，从而显著提高了渲染速度。此外，相比于存储每一帧的全连接神经网络参数，本方法通过一个共享的二维卷积神经网络动态预测全连接神经网络的参数，大大降低了存储成本。本章在 NHR 数据集[25] 和第 3 章提出的 ZJU-MoCap 数据集上评估了上述提出的渲染方法。这些数据集呈现了复杂运动的动态场景。在所有数据集上，本章提出的方法在渲染质量和速度方面表现出了最先进的水平，同时占用的存储空间较低。实验表明，本方法的渲染速度比 DyNeRF[26] 快了 100 倍。

第七章总结了全文的内容，并展望了动态人体建模与渲染领域未来的研究方向。

第 2 章 相关文献综述

动态三维人体的建模与渲染是计算机视觉与图形学领域的重要问题，已经经历了很多年的发展。为了从观测数据中创建高质量的数字人，领域内的研究人员提出了众多技术并取得了令人瞩目的成果。本章首先回顾传统的数字人建模方法，介绍这些方法的基本步骤和所需的硬件设备。其次，基于数据驱动的方法近年来发展迅速，已经成了数字人建模领域重要的技术方向。因此本章对基于数据驱动的数字人建模方法进行总结与归纳。最后，由于可微分渲染技术是目前数字人领域热门的研究思路之一，并且本文的建模技术也以可微分渲染为基础，因此本章对基于可微分渲染技术的相关研究工作进行了综述。表 2-1 列出了代表性的研究工作。

2.1 传统方法

传统的人体建模管线依赖复杂的硬件设备来采集目标人体的观测数据，如多视角图片、深度信息，从而为重建算法提供足够的观测。到目前为止，研究人员已经设计并建立了各种人体捕捉系统。这些系统通常由大量的相机组成，这些相机被精心地摆放在固定的位置，用于捕捉各个视角下目标人体的颜色和深度信息。传统的重建算法一般分为基于多视图立体匹配的重建和基于深度融合的重建。通过采集高质量的观测数据，这两类算法可以得到高精度的可渲染人体模型。本节将对这两类算法分别加以介绍。

2.1.1 基于多视角相机阵列的人体建模

多视角三维重建[95-97] 是一个广泛应用的三维重建系统。首先，该系统采集目标物体在多个相机视角下的 RGB 图片，然后使用运动恢复结构算法（Structure from motion，SfM）[96] 获得每一张 RGB 图片的相机姿态。随后，该系统使用多视图立体匹配（Multi-view stereo, MVS）[95,97] 计算每张图片的深度信息，最后使用深度融合技术得到目标三维网格模型。为了获得可渲染模型，该系统往往还会利用表面纹理贴图[98-99] 的方法，将图片中的观测映射到网格模型的纹理空间。多视角三维重建系统面临的主要挑战是难以准确恢复弱纹理区域的相机位姿和深度图，从而难以重建完整的三维网格模型。因为人体往往有弱纹理区域，比如头发和皮肤，导致基于多视角系统得到的人体模型往往不够完

表 2-1 代表性的人体建模方法。

传统方法	基于多视角相机阵列的方法	Debevec[8], Relightables[1], Collet[9]
	基于深度相机的方法	Dynamic Fusion[4]、VolumeDeform[5]、BodyFusion[6]、DoubleFusion[27]、RobustFusion[28]、POSEFusion[29]、Fusion4D[7]、Function4D[30]
基于数据驱动的方法	多边形网格	SMPLify[31]、HMR[32]、CMR[33]、Tex2Shape[34]、DeepCap[35]、METRO[36]、SPIN[37]、ROMP[38]、HybrIK[39]、PyMAF[40]、Total Capture[22]、MVPose[41]、Zhang[42]、QuickPose[43]、Kanazawa[44]、VIBE[45]、HuMoR[46]、MonoPerfCap[47]、LiveCap[48]、XNect[49]、SimPoE[50]、GLAMR[51]
	体素网格	BodyNet[52]、DeepHuman[53]、Gilbert[54]、Caliskan[55]
	隐式神经表示	PIFu[56]、PIFuHD[57]、Geo-PIFu[58]、MonoPort[59]、ARCH[60]、PAMIR[61]、ARCH++[62]、ICON[63]、Huang[64]、StereoPIFu[65]、DoubleField[66]、NHP[67]、DeepMultiCap[68]、Function4D[30]
基于可微分渲染的方法	静态场景的建模与渲染	NeRF[15]、IDR[17]、NeuS[18]、Mip-NeRF[69]、Mip-NeRF 360[70]、FastNeRF[71]、NSVF[72]、PlenOctrees[73]、KiloNeRF[74]、Plenoxels[75]、DVGO[76]、Instant-NGP[77]、TensoRF[78]、NeRFactor[79]、Zhang[80]
	动态人体的建模与渲染	Neural Volumes[81]、DyNeRF[26]、NHR[25]、Nerfies[82]、Neural Body[83]、D-NeRF[84]、Neural Actor[85]、Animatable NeRF[86]、MVP[87]、Fourier PlenOctrees[88]、Shuai[89]、HumanNeRF[90]、Relighting4D[91]、TiNeuVox[92]、InstantAvatar[93]、MPS-NeRF[94]

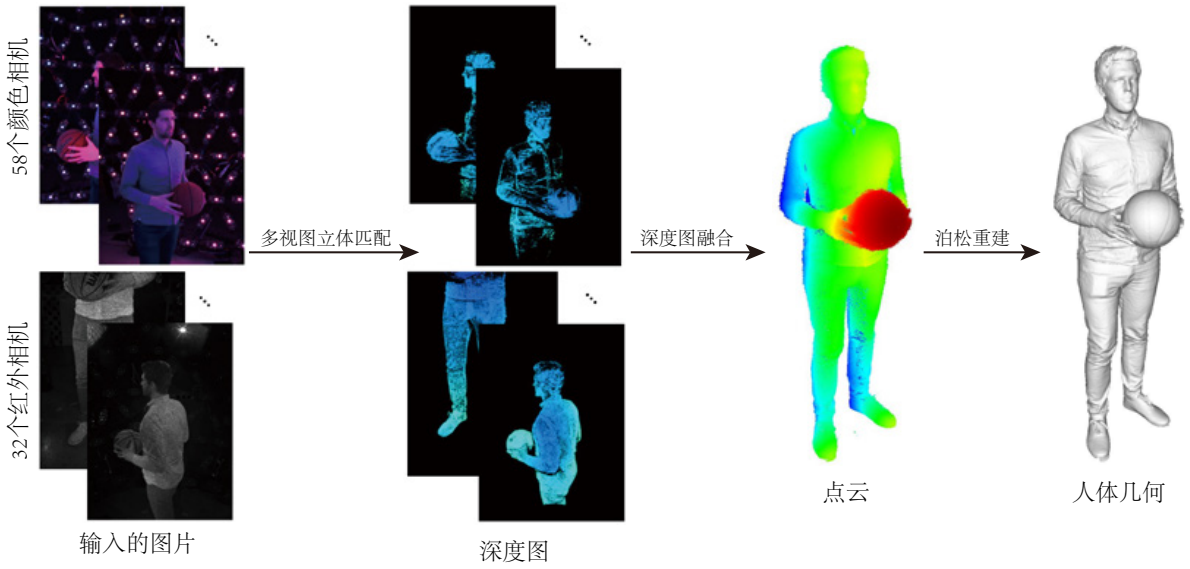


图 2-1 基于多视角立体匹配的人体建模流程[1]。

整。为了提升相机位姿估计鲁棒性和准确性，Schonberger 等人[96]提出了一种新颖的增量 SfM 算法，引入了一系列的几何验证策略、最佳视角选择、三角剖分、捆绑调整和漂移效应缓解。而在 MVS 这一步, Schonberger 等人[97]基于 PatchMatch[95] 的框架，提出了逐像素选取相机视角的策略，提升了稠密重建的效果。

除了在算法层面提升多视角重建的效果，一些方法[1,8-9] 利用复杂的硬件设备来提升重建效果。这些研究工作搭建了一个稠密视角的相机阵列，将相机摆放在固定的位置，并通过预先的相机标定[96] 获得准确的相机位姿。比如，The Relightables[1] 构建了一个大的球形穹顶，将 58 个高分辨率的 RGB 相机均匀摆放在球形穹顶。此外，该方法引入了 32 了红外摄像机，用于获得精度更高的深度图。通过深度融合技术，该方法获得了稠密的人体点云，然后通过泊松重建[100] 从点云中恢复出三维网格模型。图 2-1 展示了 The Relightables[1] 的人体重建流程。在获得高质量人体几何和外观的基础上，Debevec 等人[8]进一步地恢复了人体的材质属性，用于重光照（Relighting）等应用。该工作[8]设计了光场（Light stage），在稠密相机阵列上搭载了可编程的光源，从而实现在预设的照明条件下从多个视角捕捉目标人体的图片，然后从图片中恢复人体的材质属性。然而，该工作要求目标人体在数据采集过程中保持静止，时间约为一分钟。这样的采集过程对被拍摄者的要求较高，在实际重建中是一个比较严重的缺陷。为了捕捉动态人体，The Relightables[1] 首先恢复出每一时刻的静态人体网格模型，然后匹配不同时刻的人体网格模型，实现不同的网格模型共享一个纹理图片，从而完成从视频中恢复人体的材质参

数。为了提升材质参数的准确性，该系统使用了 331 个可编程的定制 LED 灯，增加了观测数据的光照多样性。虽然 The Relightables 可以重建出高精度的可渲染人体模型，但这个系统的弱点也比较明显。除了该系统所用的硬件设备非常复杂之外，其算法计算成本也很高，重建一个 10 秒的视频大约需要 8 个小时。

2.1.2 基于深度相机的人体建模

尽管基于多视角 RGB 相机阵列的人体重建取得了很好的效果，但此类系统漫长的计算时间无法应用于一些需要即时交互的应用中，比如远程会议、AR/VR 游戏。为了降低计算成本、实现实时重建，研究人员已经提出了一系列基于深度相机的人体建模技术[4-6,27-30,101]。作为此类方法的开创性工作，DynamicFusion[4] 实现了通过一个深度相机重建动态人体模型。具体而言，该工作使用深度相机获得目标人体每一时刻的深度图，并将深度图反投影得到世界坐标系下的点云。DynamicFusion 将第一帧的点云转为三维网格作为标准坐标系下的人体模型，而且在人体网格上定义变形图（Deformation graph）[102] 用于表示人体运动。当新的一帧点云输入时，该方法首先将点云与前一时刻的人体模型建立稠密匹配，然后基于该的匹配结果构建能量函数，随后通过最小化能量函数来优化得到变形图的参数，从而将前一时刻的人体模型变形到最新时刻，最后使用 TSDF Fusion 算法[103] 将点云融入人体模型中。通过不断地融合时序输入的深度图，该工作得到完整且精准的人体模型。然而，DynamicFusion 往往无法稳定地重建快速移动的人体。这是因为当人体快速运动时，深度相机获得的深度图与前一时刻的人体模型的偏移较大，导致难以建立准确的匹配。而且深度图可能与前一时刻的人体模型重合面积较少，以至于无法优化得到正确的变形图。

为了克服这个问题，DoubleFusion[27] 引入了人体参数化模型 SMPL[104] 用于捕捉粗糙的人体运动，并以此作为正则项约束变形图的优化过程，防止优化陷入局部最优解。虽然 DobuleFusion 实现了快速运动人体的重建，但该工作仍然无法很好地重建运动中的宽松衣物。这是因为 SMPL 模型没有建模衣物的运动，导致无法正确地正则化宽松衣物的变形图的优化过程。其他的一些研究工作[7,105] 利用硬件设备解决 DynamicFusion 遇到的问题。比如，Fusion4D[7] 架设了多个深度相机用于获得多个视角的深度图，实现基于一个时刻的观测即可重建出较为完整的人体模型，因此前后两个时刻的模型匹配和变形图估计也变得相对准确和稳定。而 Motion2Fusion[105] 使用高速深度传感器拍摄目标

人体，通过提高帧率减小了两帧之间的人体运动幅度，从而能更精确地恢复出变形图。为了简化采集设备，RobustFusion[106] 引入了数据驱动的模型重建算法，从单目 RGB-D 图片中恢复出完整的人体模型，用于帮助前后两帧人体点云的匹配与融合。该工作实现了基于单目 RGB-D 相机的动态人体高质量重建。

尽管基于深度相机的建模方法可以获得实时地重建高质量的可渲染人体模型，但深度传感器只适用于室内场景。这个问题很大程度地限制了此类重建技术的广泛应用。近年来，随着深度学习的发展，数字人领域的研究人员开始探索深度学习技术，通过数据驱动的方式学习人体几何和外观的先验，实现从更少的观测数据中重建完整的人体模型，从而减少了对采集设备的要求。

2.2 基于数据驱动的方法

为了降低人体建模的成本、提升建模系统的易用性，近年来的研究工作[53,56-57,107] 尝试使用深度学习技术从稀疏的相机视角中预测人体的几何和外观。这些研究工作表明，深度神经网络可以从三维真实数据（Ground-truth data）中学习得到三维人体模型的分布，并利用这一分布从不完整的观测（单目 RGB 图片、稀疏点云）中估计完整的人体模型。目前大多数研究工作主要采用三类的人体表示，包括多边形网格（Mesh）、体素网格（Voxel grid）、隐式神经表示（Implicit neural representation）。本节根据三维人体表示对基于数据驱动的人体建模方法进行分类。

2.2.1 多边形网格表示

多边形网格（Mesh）在计算机图形学中被广泛应用。人体建模领域中常见的多边形网格表示是参数化人体模型[104,108-109]。此类模型使用一组低维的参数向量表示三维人体，因此较为容易被神经网络拟合其分布，在单目人体重建等任务中具有很强的泛化能力。经过多年的发展，研究人员已经提出了很多种参数化人体表示，大体可分为三类：基于三角变形的模型[110-111]、基于顶点偏移的模型[104,108,112]、基于神经网络的模型[109,113-119]。这些模型通常定义了一个基准的三维网格，通过变形网格来得到不同姿态下的人体模型。基于三角变形的模型对网格三角面片施加变形函数以得到目标姿态下的人体，而基于顶点偏移的模型将网格顶点作为变形对象。一些研究工作[109,113-114] 使用深

度编解码架构回归三维网格的变形函数，从而表示更精细的人体模型。

SMPL 模型[104] 是一种基于顶点偏移的参数化人体模型，因为其适用于传统图形学的渲染流程，并且计算效率较高，所以被人体建模领域的研究工作广泛使用。具体而言，该模型是一个由形状（Shape）和姿态（Pose）参数决定的函数，其输出是一个具有 6890 个顶点的三维网格模型。SMPL 模型定义了一个基准的三维网格模型。为了表示不同形状的人体，SMPL 模型根据形状参数计算形变函数，将基准网格顶点作为变形对象，把形变应用到顶点上。当驱动基准网格模型到目标人体姿态时，SMPL 模型首先根据姿态参数计算顶点偏移量加到基准网格顶点上，用于消除变形时的网格畸变，然后再根据姿态参数计算变换矩阵，使用骨骼蒙皮驱动模型[21] 将其应用到基准网格顶点上，从而得到目标人体模型。近年来的人体建模工作[31,41,120-122] 大多致力于从输入的观测图片中预测 SMPL 模型的形状和姿态参数。

基于 SMPL 模型的方法：早期从图片中估计 SMPL 模型参数的研究工作大多会构造一个能量函数，并通过最小化这个能量函数来优化 SMPL 参数。一个经典的工作是 SMPLify[31]。这个方法首先通过利用二维卷积神经网络来定位图片中目标人体的二维关节点（2D human pose），然后将 SMPL 模型中的三维关节点投影到图片，随后计算投影关节点和预测的二维关节点的残差，通过最小化这个残差函数来优化 SMPL 参数。HuMoR[46] 使用 CVAE 模型建模了 SMPL 参数的先验，在能量函数中加入了 CVAE 模型预测的概率，从而约束 SMPL 参数的优化。Pose-NDF[120] 训练一个神经网络预测输入的 SMPL 参数到合理的 SMPL 参数的距离，作为能量函数的一部分，提升了优化的效果。虽然这些研究工作可以从单目图片中恢复出较为准确的三维人体姿态，但忽略了目标人体的面部表情和手势等细节动作。这些动作在数字人建模中非常重要，往往被人们用于传达情绪和辅助语言表达。Total Capture[22] 和 SMPL-X[108] 在 SMPL 模型的基础上加入了脸部和手的参数化模型，从而实现了包括面部表情和手势的全身人体动作捕捉。

一些研究工作[41-43] 从多视角图片中恢复 SMPL 模型参数。Dong 等人[41] 首先预测二维图片中的人体关节点，然后匹配多视角图片之间人体的关键点，随后使用三角化算法（Triangulation）将二维关节点投影到三维空间中得到三维关键点，最后拟合 SMPL 参数。为了高效地从视频中进行动作捕捉，Zhang 等人[42] 使用前一帧的多视角人体匹配关系作为当前帧的初始化，从而大大降低了计算复杂度，提升了模型推理速度。尽管基于优化的方法可以在多视角输入下产生可靠的预测，但多视角的采集设备难以推广到普

通的用户群体。而当只有单视角的图片输入时，因为从单目图片恢复三维信息存在严重的深度歧义性，这些方法的精度往往有较大的下降。此外，由于这些工作本质上是在试图解决一个复杂的非凸优化问题，所以其最终的优化收敛点很容易受到参数初始化的影响。当只存在一个视角的二维人体姿态观测时，这些工作很可能会陷入局部最优解，导致恢复出奇怪的三维人类姿态。

为了克服单目歧义性问题,近年来研究人员[32,107,122-123]探索了深度神经网络在 SMPL 参数估计这个任务中的运用。这些工作往往将图片作为神经网络的输入，然后使用神经网络直接回归 SMPL 模型的参数向量。Pavlakos 等人[107]提出了其中的一个先驱工作，首先从图片中预测二维关键点和人体分割，然后再预测 SMPL 的形状和姿态参数。为了解决网络训练中缺少真实 SMPL 参数监督的问题，该工作基于预测得到的 SMPL 参数计算相应的三维人体关节点，然后将其投影到图片空间，最后通过最小化投影点与二维关节点的重投影误差（Re-projection error）来优化模型参数。Kanazawa 等人[32]在同一时期也提出了一个端到端的单目 SMPL 参数预测方法，称为 HMR。该工作采用了对抗式训练（Adversarial training），在训练预测器的同时也训练一个判别器。此判别器用于检测网络估计的 SMPL 参数的合理性，从而最大化 SMPL 参数预测器的性能。为了从单目图片中回归合理的多人 SMPL 参数，Jiang 等人[123]设计了穿模损失函数和遮挡损失函数用于监督网络。相对于之前工作回归 SMPL 参数向量，Kolotouros 等人[33]选择预测 SMPL 网格模型的顶点位置。该工作使用二维卷积神经网络抽取输入图片的特征向量，然后将特征向量绑定到基准 SMPL 网格模型上，最好使用图卷积神经网络（Graph CNN）[124-125]提取网格顶点特征，并预测顶点的偏移量，从而变形得到目标人体模型。这个方法取得了非常好的 SMPL 模型重建效果。近期的一些研究工作[126-130]使用神经网络从多目图片中预测三维人体姿态，克服了单目歧义性问题。

当前的研究工作已经在静态人体姿态方面取得了显著的成果，近期一些研究人员开始进一步尝试从视频中估计人体运动序列。这个任务要求神经网络恢复出时序上平滑且自然的人体运动，因此非常具有挑战性。为了解决这个问题，Kanazawa 等人[44]定义了一个时间窗口，并对时间窗口内的图片序列进行编码得到时序特征，使得网络能根据时序信息来预测人体运动。该工作还从时序特征中预测了相邻帧人体姿态参数的残差，用于训练网络的时序预测能力。除此之外，该工作提出了一个单目图片特征提取器，并训练该网络使其模拟从时序图片序列中得到的时序特征，从而实现了基于单目图片估计时

序一致的人体姿态。受到 HMR[32] 的启发，Kocabas 等人[45] 提出了用于时序运动预测的对抗式训练框架 VIBE。因为判别器提供了额外的监督信号，该工作可以从非受限的环境下拍摄的视频中预测准确而自然的人体运动。

虽然人体参数化模型的低维表示使其较为容易被神经网络从稀疏图片中估计，但也使其难以表示精细的人体几何，比如面部表情和衣物细节。针对于这个问题，一些方法[34,131-132] 尝试在 SMPL 网格模型的基础上添加几何细节。给定一段单目 RGB 视频，Alldieck 等人[133] 提出一个三阶段的做法来创建可渲染的人体模型。该方法首先基于 SMPLify[31] 在时序观测上联合优化得到每一帧的三维人体姿态参数，然后基于轮廓误差来变形 SMPL 模型的网格顶点，从而建模几何细节，最后通过纹理映射技术获得网格顶点的颜色。为了使用神经网络变形模板网格模型，Zhu 等人[132] 创建了一个数据集，其中包含数千个篮球运动员的人体模型，每个模型都有高度精细的几何和纹理。基于这个数据集，该工作训练网络从单目图片预测模板人体模型的顶点偏移量，从而变形模板模型得到精细化的人体几何。考虑到在三维空间进行网络预测具有更高的复杂度，Tex2Shape[34] 提出了在二维空间预测网格变形的方法。该工作首先对输入图片预测一个 SMPL 模型，然后将图片映射到 SMPL 模型的二维纹理空间（Texture space），得到一个不完整的纹理图。基于这个纹理图，Tex2Shape 使用二维卷积神经网络在二维纹理空间预测一张位移图，然后借助 SMPL 网格与纹理空间的映射关系变形 SMPL 的网格模型。

基于模板人体网格模型的方法：此类方法[35,47] 首先建模精细的可驱动人体网格模型，然后从输入的图片中估计驱动参数，以此完成基于图片的人体重建。Xu 等人[47] 提出的 MonoPerfCap 是此类方法的代表性工作。该工作首先对目标人体扫描得到一个基准姿态下的人体模板模型，然后对其进行骨骼绑定和蒙皮，从而实现基于人体姿态参数的驱动。当输入任意图片时，MonoPerfCap 使用二维人体姿态检测器[134-135] 定位图片中的人体关节点，然后通过最小化三维人体骨骼模型到二维关节点的重投影误差来恢复三维人体姿态参数。基于此工作，Habermann 等人[48] 使用 RGB-D 图片作为输入，通过增加深度图的观测来提升人体姿态参数的稳定性，并加速了优化过程，实现了实时单目人体重建。DeepCap[35] 引入了神经网络来从单目 RGB 图片中预测人体模板的姿态参数。针对于缺少训练数据的问题，DeepCap 使用图片观测进行网络的弱监督训练，包括基于关节点的重投影误差、轮廓误差。虽然基于人体模板的方法[35,48] 可以实时地从稀疏图片中恢复高质量的人体模型，但是获取可驱动人体模型的成本较高且操作不便。

2.2.2 体素网格表示

类似于图片的像素，体素（Voxel）通常用于存放三维空间坐标的属性，比如颜色、几何、语义。与多边形网格（Mesh）相比，体素网格（Voxel grid）的数据结构比较规则，可以容易地被整合到流行的深度学习框架中。因此，很多研究工作[52-54] 尝试使用体素网格表示三维人体，并通过神经网络从稀疏图片中预测基于体素网格的人体模型。

Varol 等人[52] 提出了 BodyNet 进行单目体素网格预测。为了降低单目三维预测的歧义性，该方法采用了三阶段预测的方式，首先检测图片中目标人体的二维关键点，然后预测人体的三维关键点，最后回归人体的体素网格表示。为了训练 BodyNet，该工作定义了多视图重投影损失，将预测的人体三维体素网格投影到二维图片，然后计算与二维人体掩模的误差。Gilbert 等人[54] 利用可视外壳（Visual hull）简化三维体素网格的预测。具体而言，该工作首先多视角的二维人体掩模建立三维可视外壳，然后使用基于三维卷积层的编解码网络细化可视外壳，得到精细化的人体体素网格。DeepHuman[53] 借助了 SMPL 模型的先验来实现准确的单目人体几何估计。该工作将 SMPL 模型体素化为一个三维体素网格，然后使用基于三维卷积神经网络预测目标人体几何。虽然三维体素网络容易被网络处理，但该表示往往会占用过多的显存，限制了对高精度人体几何的建模。

2.2.3 隐式神经表示

多边形网格和体素在空间中显式地定义了三维人体的参数，而隐式神经表示（Implicit neural representation）使用一个神经网络隐式地编码空间中任意三维点的属性，比如颜色[15]、体素密度（Voxel density）[15]、符号距离（Signed distance）[14]、占据值（Occupancy value）[12]。相比于多边形网格和体素，隐式神经表示有三方面的优势。首先，因为隐式神经表示以任意空间三维点为输入，所以该表示可以建模任意分辨率的三维场景，为高质量的几何建模和图片渲染提供了可能性。其次，隐式神经表示本身由神经网络组成，因此易于和现有的深度学习技术相结合。最后，因为隐式神经表示天然可微，基于此表示构建渲染器[15-16] 可以有效地从图片中优化三维场景表示。近年来基于隐式神经表示的研究工作[56-57,64,136-139] 已经展示了很好的人体重建效果。

作为最早的使用隐式神经表示的工作之一，Huang 等人[64] 提出的人体重建方法将少量几张图片作为输入，使用二维卷积神经网络为每张图片分别提取一组多尺度特征图

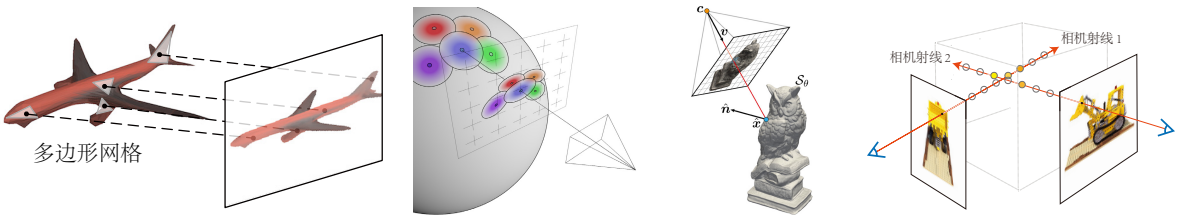
（Feature map）。基于这组多尺度特征图，该方法为任意三维点赋予一个高维特征向量，然后使用一个多层感知机（MLP）从这个特征向量中预测这个三维点的标签。如果一个三维点处于目标物体内部，那么标签为（1，0）；如果三维点在物体表面，那么标签为（1，1）；如果三维点在物体外部，那么标签为（0，1）。PIFu[56] 进一步实现了从单目图片回归三维人体几何和纹理。该工作从图片中提取特征图后，将三维点投影到二维图片，然后索引对应的图片特征向量，随后将三维点的深度与图片特征向量相叠加，再输入到一个 MLP 网络以预测占据值。当三维点在物体表面及以内，占据值为 1；当三维点在物体外部，占据值为 0。PIFuHD[57] 提出了双层几何预测器以实现高分辨率的人体重建。具体而言，PIFuHD 先使用一个 PIFu 从低分辨率的图片特征中预测人体几何，从而让网络编码的特征包含全局的人体结构。然后该工作再从高分辨率图片中提取特征图，得到三维点的特征向量后将其与第一层 PIFu 输出的特征相结合，最后用一个 MLP 网络预测人体几何。为了进一步提升重建质量，Geo-PIFu[58] 使用三维卷积神经网络从图片中提出特征体，从而让网络编码得到的特征具有三维结构，以此帮助三维信息的预测。类似地，Chibane 等人[140] 借助三维卷积神经网络从稀疏点云中提取特征体以恢复完整的人体模型。针对于多视角重建问题，StereoPIFu[65] 将 PIFu 与立体视觉结合了起来。该工作基于多视角图片构建了代价体（Cost volume），从而编码了几何信息，提升了人体模型的预测效果。ICON[63] 和 PAMIR[61] 进一步利用 SMPL 模型提升了几何预测精度。

基于隐式神经表示的数据驱动方法[56-57,64] 取得了很好的人体重建效果，然而因为该表示使用深度神经网络表示三维人体，所以不容易被显式地操控以实现人体模型的可驱动性。相比之下，参数化人体模型具有很好的可控性，但是该表示只能建模比较粗糙的人体。为了获得具有可驱动性且精细的人体模型，一些研究工作[60,62,141-143] 探索了这两种人体表示方法的结合。给定一张单目图片，ARCH[60] 首先预测图片中的 SMPL 参数，然而用骨骼蒙皮驱动算法将观测坐标系下的三维点转到 SMPL 模型定义的基准坐标系，然后在基准坐标系下回归三维人体模型。因此该工作重建得到的隐式人体模型自然地与基准坐标系下的 SMPL 模型对齐，从而实现通过 SMPL 模型驱动隐式人体模型。为了提升单目人体重建效果，ARCH++[62] 进一步利用了 SMPL 模型。该方法使用点云处理网络 PointNet++[144] 从 SMPL 模型的几何中提取为一组特征向量，用于编码人体的几何信息，然后输入到 MLP 网络中以估计三维空间的占据值。IP-Net[141] 研究了从点云中恢复可驱动的人体模型。为了实现重建的人体与 SMPL 模型的精准对齐，该工作使用

MLP 网络为每个三维点预测一个几何标签和语义标签。几何标签定义了三种空间位置下的三维点：身体内部、身体与衣物之间、人体模型外部。而语义标签编码了三维点归属的身体部位，比如头、手臂、腿。基于预测的几何标签和语义标签，IP-Net 构造了人体模型与 SMPL 模型相互匹配的能量函数，通过最小化这个能量函数以优化 SMPL 参数，实现 SMPL 模型与三维人体的对齐，从而构造出由 SMPL 模型驱动的三维人体。虽然 SMPL 模型可以实现较好的可驱动性，但 SMPL 模型中的蒙皮权重是面向一般化人体设计的，对于特定人体的驱动往往不够精细。针对于这个问题，Yang 等人借助神经网络学习了骨骼蒙皮驱动模型的先验，在重建人体模型时，除了预测人体占据值，还额外回归了三维点的蒙皮权重和骨架，从而实现了定制化的驱动模型。

虽然隐式神经表示可以实现高质量的三维人体建模，然而其重建速度通常较慢。这是因为预测一个三维点的属性需要经过一次 MLP 网络的推理，所以回归高分辨率的三维人体需要进行非常大量的网络推理。针对于这个问题，MonoPort[59] 提出了一种人体表面定位算法。该算法首先恢复一个低分辨率的三维体素网格，计算每个体素的占据值，然后再划分高分辨率的体素网格，并基于低分辨率的占据值确定人体几何可能处于的体素，随后基于 MLP 网络评估这些体素的占据值。通过这种渐进式的表面定位算法，MonoPort 省去了在大量三维点上的网络推理，减少了计算量，从而大大提升了几何重建速度。除此之外，该工作利用了几何渲染的特性进一步提升渲染速度。具体而言，当观看人体几何时，三维人体只有一面可见。因此 MonoPort 只用 MLP 网络预测了可见的三维点，从而进一步减少了网络的推理次数，实现了实时的几何渲染。NeuralHumanFVV[145] 基于 MonoPort 的渲染框架加入了人体纹理的推测，实现了实时的自由视角视频。为了完成高质量的图片渲染，该工作将多视角图片作为输入，通过图片融合的方法（Image blending）预测人体的三维纹理。FOF[146] 在减少网络推理次数的基础上提出了一个新的隐式神经表示来减少网络的推理成本。该工作使用二维卷积神经网络从图片中回归得到一张二维网格，每个网格像素中包含了一组傅立叶函数的系数。FOF 使用这组系数得到一个傅立叶函数，用于预测三维点的占据值。因为傅立叶函数的参数少于 PIFu 使用的 MLP 网络参数，所以该工作减少了预测占据值的计算量，从而提升了人体重建速度。

基于数据驱动的人体建模方法的一个主要缺陷是需要大量真实的三维数据进行深度神经网络的训练，而获得真实的三维人体数据的成本很高。之前的一些方法[56-57,60] 在合成数据集上训练人体重建网络，虽然取得了不错的效果，但在一些复杂的真实数据上



(a) 多边形网格渲染 (b) 点云渲染 (c) 隐式表面渲染 (d) 体积渲染

图 2-2 常见的渲染技术。图片分别来自 SoftRas[153]、DSS[154]、IDR[150]和 NeRF[15]。

泛化能力有限。为了解决缺少数据的问题，一些研究工作[66-67,147]开始探索可微分渲染技术，直接将观测图片作为监督信号用于训练网络，从而减少对三维数据的依赖。因为隐式神经表示可以建模连续的三维场景并且天然可微，近期的研究工作将该表示与可微分渲染进行结合得到了很好的效果。

2.3 基于可微分渲染的方法

渲染（Rendering）是将三维场景模型投影为二维图片的过程，因此从观测图片恢复三维场景模型可以被表述为一个逆向渲染问题。作为计算机图形学领域的核心问题之一，渲染技术经过了数十年的研究，已经发展成了一条较为清晰的管线，并且在电影、游戏等行业有着广泛的应用。传统渲染过程会显式地考虑很多环境变量，包括场景几何、材质属性、环境光照和相机参数。近年来，一些研究工作[148-149]定义了渲染过程中每个步骤的回传梯度，将渲染过程变得可微，然后计算渲染图片与观测图片之间的残差，通过最小化渲染误差来优化环境变量，实现了从观测图片到三维模型的逆向渲染。最近的一些工作[13,15,17,150]进一步用深度神经网络简化了渲染过程，使用网络直接预测渲染过程中的某些中间变量，以此加速渲染和提升渲染效果，并且让优化环境变量的过程变得更为稳定。这些使用神经网络进行渲染的技术被称为神经渲染（Neural rendering）[151-152]。本节将回顾近年来人体建模领域常用的可微分渲染技术，并讨论相关的人体建模工作。

2.3.1 可微分渲染技术

图 2-2 展示了多边形网格渲染、点云渲染、隐式表面渲染和体积渲染。早期方法通常使用可微分多边形网格渲染器（Mesh renderer）[148,153,155-158]。OpenDR[156]是最早的可微分网格渲染器之一，使用自动微分技术（Automatic differentiation）计算渲染过程的梯

度。Kato 等人[157] 手动定义了网格渲染的梯度函数，用于提升后向梯度计算的准确性。Neural Texture[159]、ANR[160]、MobileNeRF[161] 结合了二维卷积神经网络和网格渲染器来合成图片。Neural Texture[159] 在纹理空间定义了可学习特征图，基于三维网格将特征图映射到图像空间，然后使用二维卷积神经网络从特征图中回归生成图片。

Wang 等人[154] 提出了一个研究工作 DSS，用于可微分的点云渲染（Point cloud rendering）。该工作基于图像观测优化三维点的位置和法向量，并且考虑了遮挡点和可见性的变化。Pulsar[162] 考虑了三维点的大小，将其定义为球体，并通过可微分渲染优化球体半径。近期的一些研究工作[25,163] 在点云上绑定了高维的特征向量，并将其投影到图像空间得到二维特征图，最后基于二维卷积神经网络渲染目标图片。

近年来，一些研究工作设计了面向隐式神经表示的可微分渲染器[15-18,150,164]，用于从图片观测中优化隐式神经表示。DIST[16] 提出了面向神经隐式表面的渲染器，将球体追踪变得可微。IDR[150] 在预测神经符号距离场的同时预测了颜色场。为了渲染图片，该工作首先发射每个像素的相机射线，然后使用可微分的球体追踪求得表面点，最后预测表面点的颜色。通过最小化渲染误差，IDR 从图片中优化得到了高质量的场景几何。NeRF[15] 使用可微分的体积渲染技术来优化神经辐射场。为了渲染特定视角下的图片，该工作沿着像素的相机射线采样 N 个三维点 $\{\mathbf{x_k}\}_{k=1}^{N}$，然后预测每个采样点的颜色 $\mathbf{c}(\mathbf{x}_k)$ 和体素密度 $\sigma(\mathbf{x}_k)$。像素的颜色由体积渲染积分方程[165] 近似计算得到：

$$C(\mathbf{r}) = \sum_{k=1}^{N} T_k(1 - \exp(-\sigma(\mathbf{x}_k)\delta_k))\mathbf{c}(\mathbf{x}_k), \tag{2-1}$$

其中 $\mathbf{r}$ 是像素的相机射线，$T_k = \exp(-\sum_{j=1}^{k-1}\sigma(\mathbf{x}_j)\delta_j)$。$\delta_k = ||\mathbf{x}_{k+1} - \mathbf{x}_k||_2$ 是相邻采样点的距离。NeRF 通过最小化渲染颜色和真值颜色的误差来训练神经辐射场。VolSDF[17] 和 NeuS[18] 提出了面向符号距离场的可微分体积渲染，使得基于图片优化符号距离场的过程更为稳定，进一步提升了几何重建质量。

2.3.2 静态场景的建模与渲染

因为体积渲染器能有效稳定地基于图片优化神经三维场景表示，近几年大量研究工作[15,17-18,81,166-167] 设计了适用于体积渲染的神经场景表示，并取得了非常可观的场景建模与渲染质量。作为此类工作的先驱者，Stereo Magnification 提出了多平面图片（Multi-plane image，MPI），在场景的不同深度层（Depth plane）上分别定义一张图片。该图

片的每个像素存放颜色值和体素密度，因此可以通过体积渲染得到特定视角下的图片。虽然该工作取得了较好的渲染效果，但平面图片的几何特性限制了渲染范围。Neural Volumes[81] 提出将三维场景表示为一个三维体素网格，每个体素中存放着颜色和体素密度。三维体素网格自然地支持了 360 度的新视角渲染。然而，高分辨率的三维体素网格容易消耗大量的显存。为了解决这个问题，Neural Volumes 提出一个变换函数，使得一个低分辨率的三维体素网格可以表示高分辨率的三维场景。

为了解决三维体素网格容易占据大量显存的问题，神经辐射场 NeRF[15] 提出使用基于隐式神经表示的连续体素密度场和颜色场来建模三维场景。具体而言，NeRF 使用一个 MLP 网络预测空间中任意三维点的体素密度和颜色，因此可以表示任意分辨率的三维场景。NeRF 还提出了位置编码函数（Positional encoding），将输入的三维坐标映射为一个高维向量，从而让 MLP 网络能较好地拟合高频信号，实现了高质量的渲染。在实现体积渲染的时候，NeRF 沿着一条相机射线采样三维点，然后对这些三维点的体素密度和颜色进行积分得到像素颜色。Mip-NeRF[69] 指出这样的积分方式在渲染与训练尺度相差较远的图片时容易出现锯齿现象（Aliasing）。针对这个问题，Mip-NeRF 设计了一种新颖的渲染方式。对于每一个图片像素，该工作发射出一个三维圆锥，然后在此区域内进行神经辐射场的积分得到像素颜色。虽然 NeRF 和 Mip-NeRF 在小尺度场景渲染上取得了很好的效果，但是因为这些工作使用 MLP 网络编码场景中所有三维点的信息，导致在拟合大尺度场景时网络的容量往往不够，使得渲染质量下降。增加 MLP 网络的参加量可以解决这个问题，但也会增加计算量和神经辐射场的训练时间。Mip-NeRF 360[70] 提出了一个采样网络，首先使用该网络定位包含场景内容的三维区域，然后再用 MLP 网络预测这些区域的神经辐射场。这个策略实现了通过较小的 MLP 网络建模较大尺度的三维场景，从而实现了室外场景的高效且高质量的渲染。

NeRF 这一类研究工作[15,69-70] 让用户可以通过较为便捷的方式建模高质量的可渲染人体模型。对于一个目标人体，用户只需让其保持静止，然后用单目摄像头围绕其拍摄一圈图片，即可基于这些图片优化得到人体的神经辐射场。虽然 NeRF 在静态场景的建模与渲染上取得了非常惊人的效果，但其仍然存在一些缺陷。（1）NeRF 的渲染过程需要经过大量次数的网络推理，导致渲染速度较慢。（2）为了从多视角图片得到高质量的三维场景模型，NeRF 需要经历数小时以上的优化过程，导致计算成本较高，而且降低了用户的建模体验。（3）虽然 NeRF 能渲染高真实感的图片，然而其重建得到的场景几

何模型往往较为粗糙。（4）当前 NeRF 使用 MLP 网络预测三维点的颜色，是一个隐式的预测过程，导致该工作无法根据输入的光照显式地改变场景的外观。（5）NeRF 要求稠密视角的图片作为输入，因此限制了应用场景。近几年计算机视觉与图形学领域的研究人员提出了诸多工作来改进 NeRF 上述的缺陷。

针对于 NeRF 渲染速度慢的问题，研究工作大体从两个方面进行改进，分别是降低网络推理成本和减少网络推理次数。FastNeRF[71] 预先计算网络的输出并存放在一个离散数组中。因为不再需要进行网络推理，只需要从数组中获取相应的数值，所以 FastNeRF 大大提升了渲染速度。然而，缓存 NeRF 这个五维函数需要一个五维的数组，这容易造成非常大的存储成本。为了解决这个问题，FastNeRF 将 NeRF 拆分为两个函数。一个函数读入三维坐标并预测一组颜色向量，而另一个函数读入视角方向并预测一组权重。这组颜色向量根据权重进行加权求和得到最终的三维点颜色。通过这个策略，FastNeRF 只需要存储两个三维数组，从而降低了存储成本。相比于用一个大的 MLP 网络建模整个三维场景，DeRF[168] 和 KiloNeRF[74] 将三维空间划分为一组小区域，并分别用一个小的 MLP 网络建模这些三维区域，因此降低了网络的推理计算量。NSVF[72]、PlenOctrees[73]、SNeRG[169-170]、EfficientNeRF[171] 利用稀疏网格体素（Sparse voxels）存放神经辐射场，减少了网络的推理次数。DONeRF[172]、ENeRF[173]、AdaNeRF[174]、NeRF in detail[175]、NeuSample[176]、DDNeRF[177] 利用采样网络定位场景内容所在的三维区域，然后在此区域采样少数三维点，最后计算神经辐射场并积分得到像素颜色。AutoInt[178] 和 DIVeR[179] 利用神经网络近似体积渲染的积分过程，从而减少了相机射线上采样点的数量。基于光场技术（Light field）的研究工作[180-181] 将相机位置和相机射线方向作为神经网络的输入，直接预测像素的颜色，跳过了采样三维点的过程。MobileNeRF[161] 借助三角网格的实时光栅化（Rasterization）定位目标场景的表面几何，在二维图片空间预测最终的像素颜色，实现了在手机上的实时渲染。

为了减少 NeRF 的训练时间，研究人员已经提出了一些新的三维场景表示。考虑到神经网络优化速度较慢，Plenoxels[75] 没有使用神经网络，而是用一个离散的稀疏网格表示辐射场，从而大大加速了优化速度。DVGO[76] 定义了一个离散的三维特征体（3D feature volume），通过线性插值的方式为任意三维点赋予一个高维的特征向量，然后输入到一个小型 MLP 网络以预测辐射场。因为 MLP 网络很小，所以 DVGO 的训练速度也远远快于 NeRF。在使用小型 MLP 网络的基础上，Instant NGP[77] 提出了多尺度哈

希表（Multiresolution hash table）来将三维点映射为高维特征向量，进一步减少了训练时间。虽然这些研究工作取得了很好的效果，但离散的场景表示带来了较大的存储成本。TensoRF[78] 基于张量分解（Tensor decomposition）将三维场景解耦为一组二维平面和一组高维向量，从而减小了模型的空间复杂度。一些研究工作也探索了通过网络的预训练来减少训练时间。MetaNeRF[182] 利用元学习（Meta learning）[183] 在大量数据上训练 NeRF，得到了较好的 NeRF 初始参数，使其能较快地收敛。SRN[13]、MetaAvatar[184]、Trans-INR[185]等研究工作基于超网络技术（Hypernetwork）[19,186-188] 用一个神经网络记录多个不同的三维场景。基于多视图立体匹配，SRF[189]、IBRNet[190]、MVSNeRF[191]、NeuRay[192]、GeoNeRF[193]、PVA[194] 等研究工作使用卷积神经网络从多视角图片中提取出多张特征图，再使用一个 MLP 网络根据这些特征图预测辐射场。通过在大量数据上预训练卷积神经网络和 MLP 网络，这些研究工作具有一定的泛化能力，可以从新的多视角图片中预测出质量较高的辐射场。而且这些模型可以在新的场景图片上进一步优化，并快速收敛得到高质量的辐射场。在这些工作的基础上，DD-NeRF[147] 引入了 SMPL 模型作为人体先验，提升了辐射场预测的准确性。

针对于 NeRF 重建几何质量较差的问题，NeuS[18] 和 VolSDF[17] 使用符号距离场表示场景几何，并设计了面向符号距离场的体积渲染器，实现了从图片中恢复出高质量的三维几何模型。为了实现通过调整光照显式地改变场景外观，一些研究工作[79-80,195-197] 将隐式神经表示与基于物理的渲染（Physically-based rendering）相结合。具体而言，这些工作先定义一些 MLP 网络用于预测三维点的材质参数和几何，然后基于物理渲染模型[198] 计算三维点的颜色，最后根据体积渲染或表面渲染得到像素颜色。为了解决 NeRF 依赖于稠密视角输入的问题，一些方法[199-201] 在大量数据上训练网络，使其学习数据先验，实现从稀疏视角图片中预测较好的隐式场景表示。pixelNeRF[199] 通过体积渲染从多视角图片中学习辐射场预测网络的先验，从而减少了对真实三维数据的依赖。一些研究工作[202-207] 使用对抗式训练框架实现了从单目图片数据集中学习三维数据的分布，进一步弱化了对训练数据的要求。

2.3.3 动态人体的建模与渲染

最近，一些研究工作开始尝试使用可微分渲染器建模动态场景，实现了低成本地创建动态数字人。对于目标动态场景，Neural Animated Mesh[3] 在每一个时刻分别用一个

MLP 网络编码该时刻下的三维场景。虽然该策略可以完成高质量的可渲染场景建模，然而也导致了较大的训练成本和存储成本。一些方法探索了通过一个神经网络表示一个动态场景。Neural Volumes[81] 使用一个三维卷积神经网络预测不同时刻下的三维体素网格，用于存放颜色和体素密度。为了建模连续的动态场景，DyNeRF[26] 将时间的隐变量作为 NeRF 额外的网络输入，从而可以表示不同时刻的三维场景，也即构建了动态的三维场景。NHR[25] 借助了显式的三维点云序列来表示动态场景。该工作使用 PointNet++[144] 从三维点云提取高维特征向量，然后将点云特征投影为目标视角下的特征图，最后用二维神经网络渲染得到目标图片。虽然这些研究工作可以从输入的多视角视频中重建高真实感的动态人体自由视角视频，然而仍然存在一些缺陷。首先，这些工作通常要求稠密视角的视频作为输入，导致其依赖于复杂的硬件设备来采集数据，从而提升了建模成本。其次，重建得到的动态数字人往往无法被显式地驱动，限制了应用场景。最后，上述工作的渲染速度较慢，因此用户无法实时地切换观看视角。

为了从稀疏视角视频中重建自由视角视频，Neural Body[83] 将参数化人体模型模型（SMPL）与神经辐射场相结合，利用人体先验整合了不同视频帧的观测信息。具体而言，Neural Body 在 SMPL 模型的网格顶点上绑定一组可学习隐变量，通过变化这组隐变量的空间位置来表示不同时刻下的数字人体。对于一个特定的视频帧，该工作使用三维卷积神经网络[208-209] 将这组隐变量转换为神经辐射场。因此，Neural Body 可以从同一组隐变量中恢复每一帧的三维人体模型，从而自然地整合了时序观测，实现了稀疏视角下的人体重建。最近的一些研究工作[82,84-85,90,210-214] 将动态场景分解为一个标准静态场景模型和一个变形场，通过变形场将不同时刻的信息显式地整合到静态场景模型中，因此也能从稀疏视角视频中重建三维动态人体。HumanNeRF[90] 表明了此类做法可以从单目视频中恢复出高质量的人体模型，超过了 Neural Body 的渲染质量。

可驱动的人体模型对于各种数字人应用是必不可少的，例如游戏、沉浸式虚拟会议、虚拟伴侣。ANR[215]、SMPLpix[216] 等工作借助 SMPL 模型得到目标人体姿态和目标视角下的特征图，然后使用二维卷积神经网络预测目标图片。然而，一些研究工作[83,86] 表明二维卷积神经网络难以保证视角之间渲染的连续性。最近的一些方法[85-86,217-218] 用一个标准坐标系下的神经辐射场和一个变形场来表示动态人体，其中变形场用于建立观察空间和规范空间之间的对应关系。Animatable NeRF[86] 是其中的代表性工作。为了实现数字人的可驱动性，Animatable NeRF 基于骨骼蒙皮驱动模型，将人体姿态参数与蒙皮

权重权重场相结合来表示变形场。针对可驱动数字人的渲染质量问题，一些工作[219-221]首先建模一个高精度的人体模型，然后再通过骨骼蒙皮驱动算法操纵人体模型，最后基于神经渲染得到目标图片。

针对于动态数字人体的实时渲染问题，研究人员基于神经辐射场提出了各种先进的动态场景表示，以减少渲染所需的计算量。MVP[87]、Drivatar[222] 将动态场景表示为一个多边形网格序列，并在多边形网格表面绑定一组离散三维网格体素，用于存储更精细的场景几何和纹理，最后用体积渲染合成图片。这样的策略大大减少了三维空间中的采样点数量，并且无需推理 MLP 网络，进一步降低了计算成本，实现了实时渲染。然而，MVP 和 Drivatar 所需的多边形网格序列难以获取，限制了他们的使用。FastNeRF[71] 将动态场景表示为一个标准坐标系下的神经辐射场和一个变形场。该方法预先计算了神经辐射场并使用三维网格体素进行存储，因此提升了整个模型的推理速度。Fourier PlenOctrees[88] 通过傅立叶变换将动态场景表示为一个三维网格体素，其中每个体素存放相应的傅立叶系数。虽然该方法提升了渲染速度，但也大大增加了存储成本。ENeRF[173] 利用一个可泛化网络从输入的多视角图片中推理得到场景几何表面，然后在几何表面附近采样三维点，最后预测这些三维点的颜色和体素密度。通过减少采样点速度，该工作大大降低了计算成本。除此之外，ENeRF 所需的存储成本较小，因为该工作将动态场景表示为多视角视频和一个可泛化网络，而多视角视频可以被二维视频编码技术高效地压缩。然而，ENeRF 难以在较大范围的新视角下合成高真实感的图片。

近年来，一些研究工作[67,89,91-92,223-225] 在各个方向提升了动态场景模型的能力。Relighting4D[91] 拓展了 NeuralBody 的网络，额外预测了动态人体的材质参数，实现了可重光照的人体模型。Shuai 等人[89] 在 Neural Body 的基础上额外建模了静态背景，并且通过多人动作捕捉实现多人动态场景的建模。一些研究工作[92-93,224,226-227] 将动态人体表示为静态模型和变形场，并利用 Instant-NGP[77]、DVGO[76] 等技术加速静态模型的训练。针对于网络训练速度慢的问题，Neural Human Performer[67] 和 MPS-NeRF[94] 在大量数据上预训练了一个可泛化网络，实现了从输入的视频中推理得到三维动态人体模型。

第 3 章　基于结构化隐变量的人体神经辐射场表示

3.1　引言

从视频中重建数字虚拟人并制作动态人体的自由视角视频具有广泛的应用，比如电影特效、体育广播和远程虚拟会议。早期的自由视角视频系统大多依赖于稠密相机阵列拍摄目标人体的多视角图片，从而通过基于图像的视角合成[10,228]以产生逼真的渲染。另外一些研究工作通过深度相机进行人体的高质量三维重建[7,9]以生成自由视角视频。虽然这些研究工作取得了很好的效果，但其所需的复杂硬件较为昂贵且只适用于受限的环境中，使得这些自由视角视频系统难以推广。

本章主要关注从稀疏视角视频中重建出动态人体的自由视角视频。图 3-1 给出了一个例子，其中输入的稀疏视角视频由少数的同步摄像机所拍摄。这种设定大大降低了采集输入数据的难度和创作自由视角视频的成本，并且适用于更广泛的用户。然而，这个设定非常具有挑战性。传统的基于图像的渲染方法[10,229]大多需要密集的输入视图，因此无法应用于这个稀疏视角的设定下。而对于基于三维重建的方法[1,96]，其重建流程中的稠密立体匹配要求输入图片之间的相对相机位姿较小，所以也不适用于当前宽基线（Wide baseline）的情况。此外，因为人体部位通常存在自遮挡的现象，当目标人体只被少数几个相机时，人体的一部分在某一时刻的图片中很可能是不可见的。因此，传统的重建方法往往会产生嘈杂和不完整的几何结果，导致较差的渲染质量。

近年来，一些研究工作[13,15,164] 探索了隐式神经表示在新视角合成中的潜力。神经辐射场 NeRF[15] 表明，通过将三维场景表示为密度和颜色的隐式神经场，可以实现照相级的视角合成效果。为了从图片中获得神经辐射场，该工作使用可微渲染器将神经辐射

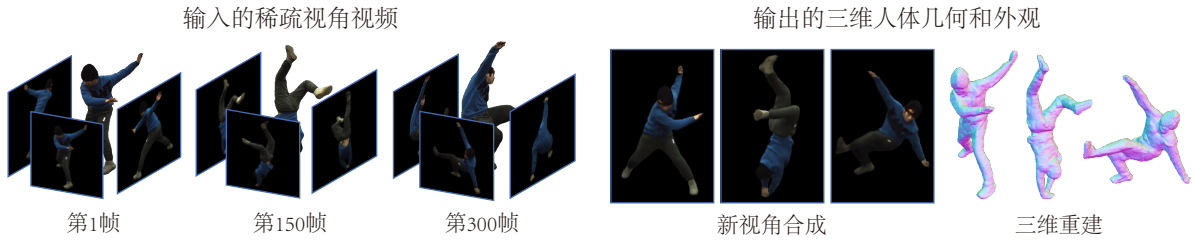


图 3-1 本章提出了一种基于结构化隐变量的动态人体隐式神经表示方法，用于从稀疏视角视频中合成自由视角视频。本方法通过可微分体积渲染在输入视频上优化人体表示，从而重建出高质量的三维人体模型，用于新视角合成和三维人体几何重建。

场渲染为图片，并和输入图像进行比对，通过最小化渲染误差以优化神经辐射场。然而，本章第 3.3.1 节中的实验结果表明，当输入图像的视角数量较低时，神经辐射场[15] 的性能会急剧下降。该现象的原因是从图片恢复三维信息存在歧义性，而非常稀疏的观测图片无法为优化过程提供足够的约束以消除歧义性，导致从图片中获得神经辐射场成为一个病态问题。解决该病态问题的一个方法是整合输入视频不同时刻的观测信息以增加对优化过程的监督信号。为了实现这个思想，Neural Volumes[81] 对于不同视频帧使用相同的神经网络预测相应的三维场景表示。然而，Neural Volumes 中的神经网络输入为一个每一视频帧独立获得的隐变量，缺乏足够的时序关联性，导致该模型无法有效地融合跨帧观测。本章第 3.3.1 节的实验结果显示 Neural Volumes 的时序重建结果较差。

针对从稀疏视角视频合成自由视角视频这一挑战，本章提出了一种新颖的基于隐式神经表示的动态人体模型。图 3-2 展示了该人体模型的基本思想。对于不同帧的人体模型，本章提出的方法不像之前的研究工作[15] 去单独学习它们，而是从一组共享的隐变量（Latent code）中回归不同帧的三维人体。具体而言，本方法预先定义一组隐变量并将其锚定到可变形人体模型（本工作中为 SMPL[104]）的网格顶点上，使得这些隐变量的空间位置可随人体姿态而变化。为了获得特定视频帧下的三维人体，本方法首先使用人体姿态检测方法[31,121,230] 从输入的稀疏视角图片中获得人体姿态参数，然后根据人体姿态转换隐变量的空间位置。最后，本章设计了一个神经网络从这组隐变量中回归任意

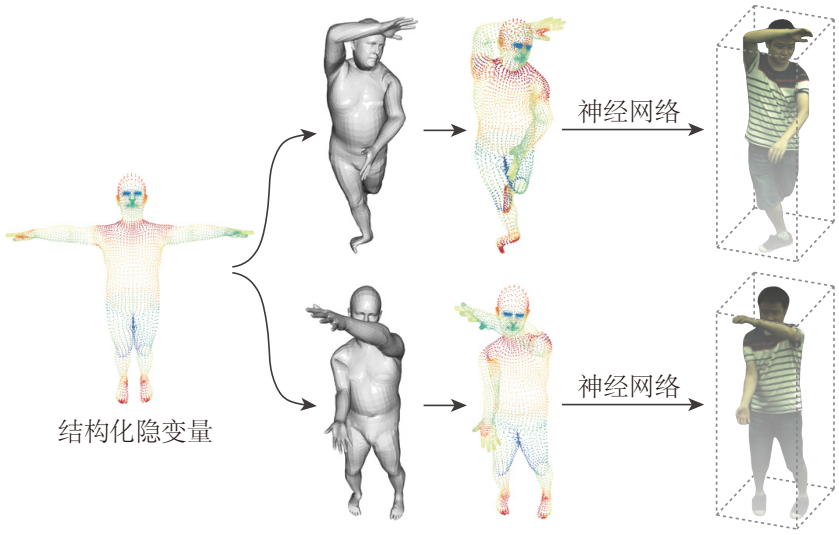


图 3-2 本方法将一组隐变量绑定到一个可变形人体网格模型的顶点上，然后从同一组隐变量中回归不同帧的三维人体表示。对于每一个视频帧，本方法基于三维人体姿态将隐变量的空间位置进行变换，然后使用一个 MLP 网络从结构化的隐变量中回归任意三维位置的体素密度和颜色。最后，本章提出的动态人体表示可以通过体积渲染合成任意视角下的图片。

三维点的体素密度和颜色。本方法通过可微分渲染在输入视频的所有视频帧上联合优化预定义的隐变量和神经网络参数，从而有效地整合不同帧的观测。本章提出的动态人体模型可以看作是统计学中的隐变量模型（Latent variable model）[231]的一种实现。该方法的另一个优点是，可变形模型粗略的几何表面为优化过程提供了隐式的几何约束，从而可以更有效地学习目标人体模型。

综上所述，本章贡献如下：首先，本章提出了一个新颖的自由视角视频系统，输入一组由少数同步相机拍摄的多视角视频，该系统可以实现高真实感的自由视角渲染，支持用户从任意视角观看采集的动态人体。其次，本章提出了一个基于隐式神经表示的动态人体模型，通过将一组隐变量绑定到可变形人体模型，实现从一组共享的隐变量中生成不同视频帧的人体模型，从而在优化过程中整合了输入视频的时序信息，解决了稀疏视角重建的病态问题。最后，本章采集了一个多视角视频数据集，在该数据集上验证了提出的方法相比于之前的技术由显著的性能提升，通过充分的消融实验证明了该方法各个模块的有效性，并探究了该方法在不同视角数量与不同长度的视频上的性能。

3.2 方法

3.2.1 方法概述

给定一个动态人体的稀疏多视角视频，本工作的目标是生成一个动态人体的自由视点视频。本工作将多视角视频表示为 $\{\mathcal{I}_t^c | c = 1, ..., N_c, t = 1, ..., N_t\}$，其中 c 是相机索引，N_c 是相机数量，t 是帧索引，N_t 是视频的帧数。本工作假设输入的多视角视频的相机参数已经被预先校准。除此之外，对于输入视频中的每张图像，本工作使用一个人体分割算法[232]来获得前景人体的掩模（Mask），并将图像背景的像素值设为零。

本工作所提出模型的概述如图 3-3 所示。具体而言，本文提出的方法首先在可变形人体模型表面上绑定一组结构化隐变量（第 3.2.2 节），然后通过隐变量扩散的过程得到三维空间中任意点的隐变量（第 3.2.3 节），最后使用一个 MLP 网络（第 3.2.4 节）将隐变量解码为体素密度和颜色值。本工作使用体积渲染投影数字人体模型得到任意视点下的图像，并通过通过最小化渲染图像和输入图像之间的差异（第 3.2.5 节）来联合优化结构化隐变量和神经网络的参数。

本工作从同一组结构化隐变量中生成不同视频帧的数字人的几何和纹理。从统计学

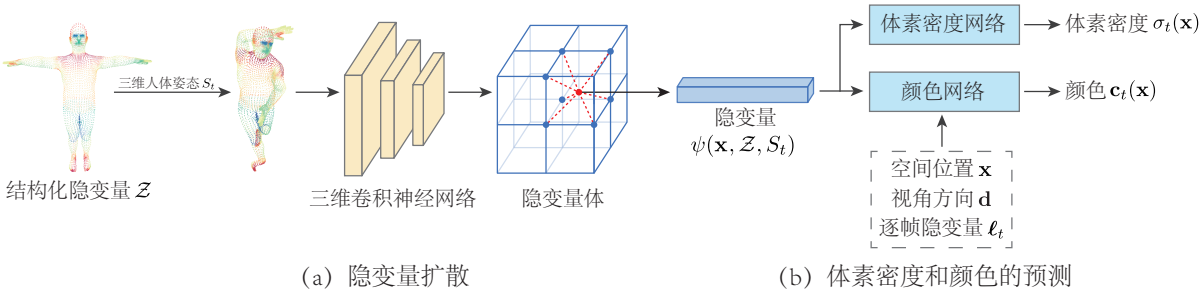


图 3-3 本工作基于结构化隐变量构建动态人体模型。（a）结构化隐变量被输入到一个三维卷积神经网络中，其输出一个隐变量体。该过程将输入的结构化隐变量扩散到邻近人体模型表面的三维空间。（b）对于三维空间中的任意点，其隐变量通过三线性插值从其相邻的体素中获得，并输入到一个 MLP 网络中进行体素密度和颜色回归。

的角度来看，这是一种隐变量模型[231]。该模型将每一视频帧的观测与同一组隐变量联系起来。基于这样的隐变量模型，本工作可以有效地整合输入视频中的观测信息。

3.2.2　结构化隐变量

为了使用人体姿态控制隐变量的空间位置，本工作将这些隐变量锚定在一个可变形人体模型（SMPL 模型）[104]上。SMPL 模型是一个三维网格模型，其被定义为形状参数、姿态参数和刚性变换矩阵参数的函数，该函数的输出是一个具有 6890 个网格顶点的三维人体模型。具体而言，本工作在 SMPL 模型的顶点上定义一组隐变量 $\{\mathcal{Z} = \boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_{6890}\}$。对于输入视频第 t 帧，本工作使用人体姿态检测器[22]从多视角图片 $\{\mathcal{I}_t^c | c = 1, ..., N_c\}$ 中估计出 SMPL 参数 S_t，然后基于人体姿态 S_t 变换隐变量的空间位置，最后从结构化隐变量中回归人体体素密度和颜色值。图 3-3 呈现了结构化隐变量的一个例子。在本章的实验中，隐变量 $\boldsymbol{z}$ 的维度被设置为 16。

类似于局部隐式表示[233-235]，隐变量与深度神经网络一起被用于表示人体局部的几何和外观。本工作将这些隐变量锚定到可变形模型上，从而能够表示一个动态数字人体。基于这个动态人体表示，本工作建立了一个隐变量模型，将一组相同的隐变量映射到不同视频帧的体素密度和颜色的隐式神经场，从而自然地整合了时序观测。

3.2.3　隐变量扩散

图 3-3（a）展示了隐变量扩散的过程。因为隐式神经场需要给三维空间中的每个点预测体素密度和颜色值，所以其需要在任意的三维空间点上查询隐变量。基于三线性插

表 3-1 三维卷积神经网络的网络层结构。每一个网络层由一个稀疏三维卷积、批量归一化（Batch normalization）和 ReLU 组成。

	网络层描述	网络输出维度
	输入特征体	D×H×W×16
1-2	(3 × 3 × 3 卷积核, 16 通道, 1 步长) × 2	D×H×W×16
3	3 × 3 × 3 卷积核, 32 通道, 2 步长	1/2D×1/2H×1/2W×32
4-5	(3 × 3 × 3 卷积核, 32 通道, 1 步长) × 2	1/2D×1/2H×1/2W×32
6	3 × 3 × 3 卷积核, 64 通道, 2 步长	1/4D×1/4H×1/4W×64
7-9	(3 × 3 × 3 卷积核, 64 通道, 1 步长) × 3	1/4D×1/4H×1/4W×64
10	3 × 3 × 3 卷积核, 128 通道, 2 步长	1/8D×1/8H×1/8W×128
11-13	(3 × 3 × 3 卷积核, 128 通道, 1 步长) × 3	1/8D×1/8H×1/8W×128
14	3 × 3 × 3 卷积核, 128 通道, 2 步长	1/16D×1/16H×1/16W×128
15-17	(3 × 3 × 3 卷积核, 128 通道, 1 步长) × 3	1/16D×1/16H×1/16W×128

值，本工作可以基于离散的隐变量获得任意空间点的隐变量。然而，由于结构化隐变量在三维空间中相对稀疏，直接插值隐变量会导致大多数三维点的隐变量为零向量。为了解决这个问题，本工作将 SMPL 模型表面上的隐变量扩散到表面附近的三维空间。

受之前研究工作[236-238] 的启发，本工作选择了 SparseConvNet[209] 来高效地处理结构化隐变量。表 3-1 中描述了这个网络的详细架构。具体来说，根据 SMPL 参数，本方法计算了人体的三维边界框，然后将边界框划分为小体素，每个体素的大小为 $5mm \times 5mm \times 5mm$。非空体素的隐变量是体素内 SMPL 顶点的隐变量平均值。SparseConvNet 使用三维稀疏卷积来处理输入的隐变量，并输出具有 $2\times, 4\times, 8\times, 16\times$ 下采样大小的隐变量体（Latent code volume）。通过三维卷积和下采样，输入的隐变量被扩散到模型表面的附近空间。本工作参考之前的研究工作[237]，对于三维空间中的任何一点，从第 5、9、13、17 个网络层得到的多尺度隐变量体中插值隐变量，并将这些隐变量连接得到最终隐变量。考虑到隐变量扩散不应该受到世界坐标系中人体位置和方向的影响，本方法将结构化隐变量预先转换到 SMPL 坐标系下。

对于三维空间中的任意点 $\mathbf{x}$，本工作从隐变量体中查询该点的隐变量。具体来说，本工作首先将三维点 $\mathbf{x}$ 转换到 SMPL 坐标系，在 3D 空间中对齐三维点和隐变量体。然后，本工作使用三线性插值计算得到该点的隐变量。对于 SMPL 参数 S_t，三维点 $\mathbf{x}$ 的

隐变量被表示为 $\psi(\mathbf{x}, \mathcal{Z}, S_t)$。三维点的隐变量最终被传入 MLP 网络中，用于预测该点的密度和颜色。

3.2.4 体素密度和颜色的预测

图 3-3（b）概述了三维空间中任意点的体素密度和颜色回归。体素密度和颜色的隐式神经场由 MLP 网络表示。

体素密度模型：对于输入视频的第 t 帧，本工作将在点 $\mathbf{x}$ 处的体素密度函数定义为：

$$\sigma_t(\mathbf{x}) = M_\sigma(\psi(\mathbf{x}, \mathcal{Z}, S_t)), \tag{3-1}$$

其中 M_σ 表示具有四层全连接层的 MLP 网络。

颜色模型：与之前的研究工作[15,81]类似, 本方法在预测颜色时同时考虑了视角方向 $\mathbf{d}$ 和隐变量 $\psi(\mathbf{x}, \mathcal{Z}, S_t)$。为了模拟入射光随空间位置变化的情况，本方法的颜色模型也将空间位置 $\mathbf{x}$ 作为输入。本工作发现不同时刻下一些环境因素会发生变化，从而影响了人体的外观。受到自动解码器（Auto-decoder）[14]的启发，本方法为每个视频帧分配了一个可学习的特征向量 $\boldsymbol{\ell}_t$ 来编码这样的时变因素。

具体来说，对于第 t 个视频帧，空间点 $\mathbf{x}$ 的颜色值被定义为关于隐变量 $\psi(\mathbf{x}, \mathcal{Z}, S_t)$、视角方向 $\mathbf{d}$、空间位置 $\mathbf{x}$、以及时变特征向量 $\boldsymbol{\ell}_t$ 的函数。为了更好地学习高频函数，本工作将位置编码[15,239] 应用于视角方向 $\mathbf{d}$ 和空间位置 $\mathbf{x}$。第 t 帧的颜色模型定义为：

$$\mathbf{c}_t(\mathbf{x}) = M_\mathbf{c}(\psi(\mathbf{x}, \mathcal{Z}, S_t), \gamma_\mathbf{d}(\mathbf{d}), \gamma_\mathbf{x}(\mathbf{x}), \boldsymbol{\ell}_t), \tag{3-2}$$

其中 $M_\mathbf{c}$ 是一个具有两层全连接层的 MLP 网络，而 $\gamma_\mathbf{d}$ 和 $\gamma_\mathbf{x}$ 分别是用于视角方向和空间坐标的位置编码函数。本章中的实验将时变特征向量 $\boldsymbol{\ell}_t$ 的维度设为 128。

3.2.5 模型训练细节

给定一个目标视角，本方法使用可微分的体积渲染技术将上文提出的动态人体表示渲染为二维图片。本工作通过最小化观察图像 $\{\mathcal{I}_t^c | c = 1, ..., N_c, t = 1, ..., N_t\}$ 的渲染误差来优化模型参数：

$$\underset{\{\boldsymbol{\ell}_t\}_{t=1}^{N_t}, \mathcal{Z}, \Theta}{\text{minimize}} \sum_{t=1}^{N_t} \sum_{c=1}^{N_c} L(\mathcal{I}_t^c, P^c; \boldsymbol{\ell}_t, \mathcal{Z}, \Theta), \tag{3-3}$$

其中 Θ 表示网络参数，P^c 是相机参数，L 是总平方误差，衡量渲染图像和观察图像之间的差异。相应的损失函数定义为：

$$L = \sum_{\mathbf{r}\in\mathcal{R}} \left\| \tilde{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2, \tag{3-4}$$

其中 $\mathcal{R}$ 是穿过图像像素的相机光线集合，$C(\mathbf{r})$ 表示真实的像素颜色。与逐帧的重建方法[15,96]相比，本文的方法使用视频中所有的图像来优化模型，从而可以利用更多的信息用于恢复三维场景结构。

本工作采用了 Adam 优化器[240]来训练模型参数。初始学习率设置为 $5e^{-4}$，并在优化过程中指数衰减到 $5e^{-5}$。本方法提出的模型在四块 2080Ti GPU 上进行训练。一个 4 视角的 300 帧视频通常需要 20 万次迭代才能收敛（大约 14 小时）。

3.2.6 应用

在训练之后，本方法提出的模型可以用于生成动态人体的自由视角视频和三维几何模型。本方法通过体积渲染合成自由视角视频，使得观众可以从任意角度观看动态人体的视频。本章的实验结果表明，本方法生成的视频表现出较高的帧间和视图间的渲染一致性。对于三维人体模型重建，本方法首先将场景离散化为一个三维体素网格，每个体素的大小为 $5mm \times 5mm \times 5mm$。然后，本方法计算所有体素的体积密度，并用 Marching Cubes 算法[241] 提取人体多边形网格模型。

3.3 实验分析

3.3.1 ZJU-MoCap 数据集上的实验

为了评估本章所提出的方法，本工作创建了一个名为 ZJU-MoCap 的多视角数据集。这个数据集使用具有 21 个同步相机的多相机系统拍摄了 9 个动态人体视频。在实验中，本工作选择四个均匀分布的相机视角进行训练，并使用剩余的相机视角进行测试。该数据集中所有视频序列的长度在 60 至 300 帧之间。数据集包含了复杂的动作，包括扭转、太极、手臂摆动、热身、拳击和踢腿。

实验指标：对于图片合成，本工作参考 NeRF[15] 使用两个标准指标：峰值信噪比 (PSNR) 和结构相似性指数 (SSIM) 来评估各个方法。对于三维重建，因为没有真实的人

表 3-2 ZJU-MoCap 数据集上的新视角合成的 PSNR 结果。“NV” 表示 Neural Volumes，“NT” 表示 Neural Textures。本方法在所有视频序列上的渲染质量都优于之前的方法。

	PSNR ↑			
	NV[81]	NT[159]	NHR[25]	本方法
Twirl	22.09	25.78	26.68	**30.56**
Taichi	18.57	19.44	19.81	**27.24**
Swing1	22.88	24.96	24.73	**29.44**
Swing2	22.08	24.84	25.01	**28.44**
Swing3	21.29	23.50	23.47	**27.58**
Warmup	21.15	23.74	23.79	**27.64**
Punch1	23.21	24.93	25.02	**28.60**
Punch2	20.74	22.44	22.88	**25.79**
Kick	22.49	24.33	23.72	**27.59**
平均	21.39	23.77	23.90	**28.10**

体三维几何形状，本章的实验只提供了定性结果。

新视角合成比较：本章将提出的模型与之前的视图合成方法[25,81,159]进行比较。所有方法都是在每个场景上单独训练一个网络。（1）Neural Volumes[81]将每个视频帧的多视角图像编码为隐变量并将其解码为离散化的 RGBα 三维体素网格。（2）Neural Textures[159]在纹理空间定义可学习的特征图，然后基于粗糙的三维网格模型将特征图映射到二维图片空间，最后用二维卷积神经网络预测为图片。由于 Neural Textures[159] 没有开源，所以本工作复现了这个研究工作。并且为了公平比较，本工作将 SMPL 网格作为 Neural Textures 的输入。（3）NHR[25] 使用深度神经网络将输入点云渲染为图像。本章将 SMPL 顶点作为 NHR 的输入点云。

表 3-2 和 3-3 展示了本方法与之前研究工作[25,81,159]的比较。在 PSNR 指标和 SSIM 指标上，本章提出的模型都在所有方法中取得了最佳性能。具体而言，本方法至少比之前的方法高出 4.20 PSNR 和 0.047 SSIM。相比于通过逐帧隐变量建模动态人体[81]，本方

表 3-3 ZJU-MoCap 数据集上的新视角合成的 SSIM 结果。本方法的渲染质量远远好于之前方法。

	SSIM ↑			
	NV[81]	NT[159]	NHR[25]	本方法
Twirl	0.831	0.929	0.935	**0.971**
Taichi	0.824	0.869	0.874	**0.962**
Swing1	0.726	0.905	0.902	**0.946**
Swing2	0.843	0.903	0.906	**0.940**
Swing3	0.842	0.896	0.894	**0.939**
Warmup	0.842	0.917	0.918	**0.951**
Punch1	0.820	0.877	0.879	**0.931**
Punch2	0.838	0.888	0.891	**0.928**
Kick	0.825	0.881	0.873	**0.926**
平均	0.821	0.896	0.897	**0.944**

法从相同的一组结构化隐变量中生成不同帧的人体隐式神经表示。结果表明，本方法更好地整合了时序信息，得到了更好的渲染结果。

图 3-4 展示了本方法和其他方法[15,25,81,159] 的定性结果。Neural Volumes[81] 的渲染结果表明，该工作不能准确重建三维人体几何形状和纹理。Neural Volumes[81] 的渲染结果较为模糊。作为图像到图像转换（Image-to-image translation）的方法，Neural Textures[159] 和 NHR[25] 难以正确地生成相应视角下的图片。相比之下，本方法可以在新视角下合成高质量的图片。

三维重建比较：本工作在收集的多视图数据集 ZJU-MoCap 上测试了多视角重建算法 COLMAP[96-97]和 DVR[164]。COLMAP[96-97] 是一个发展了很多年的多视角立体重建算法，而 DVR[164]通过可微渲染器学习占用场（Occupancy field）[12]。实验结果表明 COLMAP 和 DVR 无法从四个输入视角中恢复合理的三维人体形状。

为了进行比较实验，本工作选择了一种基于学习的预训练方法 PIFuHD[57] 作为基线方法。PIFuHD 在 450 个高精度三维人体模型上训练单视图重建网络。本工作使用其

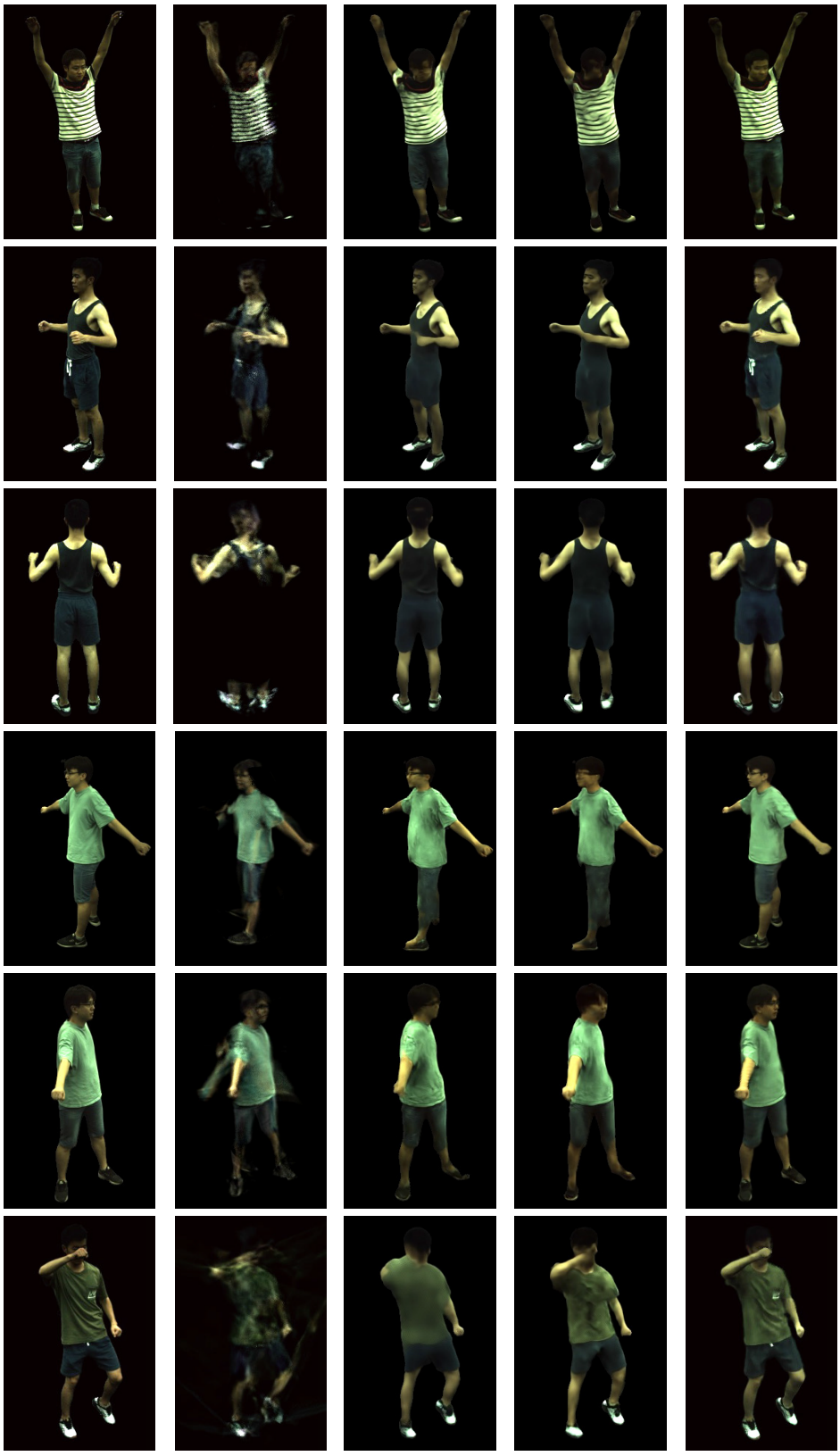


图 3-4 ZJU-MoCap 数据集上的新视角合成的定性比较。“NV” 表示 Neural Volumes[81]，“NT” 表示 Neural Textures[159]。输入视频由四个同步相机拍摄。本实验选择两个新视角进行定性比较。本章的方法明显优于之前的方法[81,159]。

图 3-5 ZJU-MoCap 数据集上的三维人体几何重建的定性比较。本章提出的模型能够高质量地重建三维人体几何模型，而且能重建较为宽松的衣服，如第三个人的连帽衫。PIFuHD[57] 难以在 ZJU-MoCap 取得合理的重建结果。

开源的代码和预先训练的模型进行推理，将多视角视频的第一个视图作为 PIFuHD 的输入。Deep volumetric video[64]和 PIFu[56] 提出了预训练的多视图重建网络，但这两个工作没有开源预先训练好的网络模型。

图 3-5 展示了本章提出的方法和 PIFuHD 之间的定性比较。实验结果表明本方法能够从稀疏视角的视频中重建精确的人体几何形状。由于本方法从多视角图片中基于可微分渲染器优化三维人体表示，所以重建出的人体模型的三维人体姿态与图片中观测得到的结果高度一致。PIFuHD 的重建结果表明，这些工作在 ZJU-MoCap 数据集上没有很好的泛化能力。对于具有复杂人体姿态的人体图片，PIFuHD 无法恢复正确的人体形状。此外，其重建模型与从多视图图像观察到的人体不一致。

3.3.2 People-Snapshot 数据集上的实验

为了验证本章提出的方法可以从单目视频中重建动态人体，本工作在单目数据集 People-Snapshot[131] 上做了实验。这个数据集拍摄了一些人体，每个人保持 A 姿态在原地转圈。因为此类动作较为简单，所以 People-Snapshot 可以从单目视频中准确估计

图 3-6 单目视频上的新视角合成效果。与 People-Snapshot[131]相比，本方法可以渲染出更多的外观细节，如图中第一个人的衬衫和第二个人的裤子。

人体的 SMPL 参数。本工作还与 People-Snapshot[131] 提出的方法进行了比较。People-Snapshot 通过变形 SMPL 模型的顶点来拟合输入视频中的二维人体轮廓。根据 People-Snapshot[131]的实验设置，本工作只在该数据集上进行了可视化比较。

新视角合成比较：图 3-6 展示了新视角合成的定性比较。与 People-Snapshot[131] 提出的方法相比，本章提出的模型可以渲染更多的纹理细节，并且可以更好地重建穿着宽松衣物的目标人体。如图 3-6 所示，本模型可以准确地建模图中第一个人的上衣，而

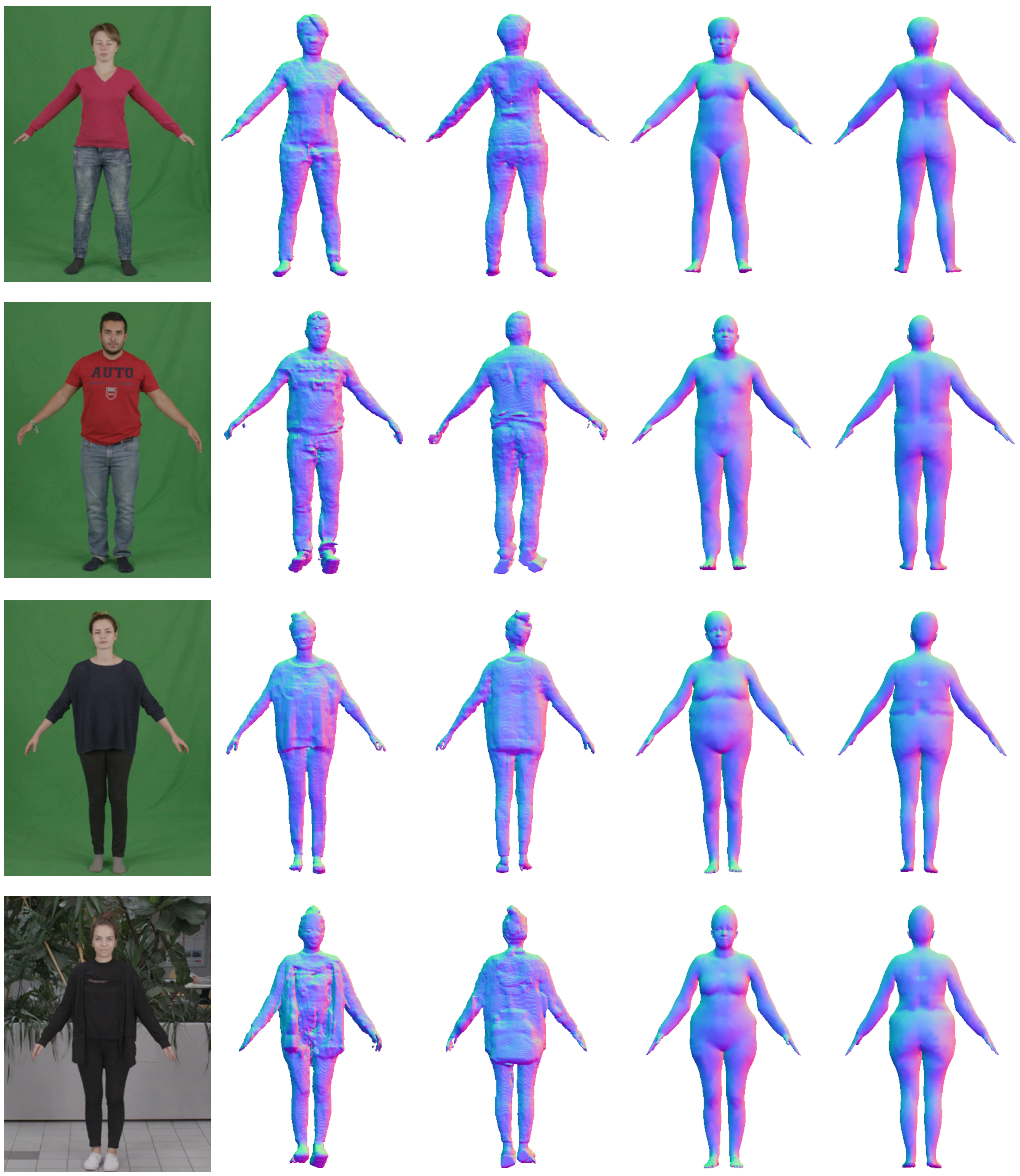


图 3-7 单目视频上的三维几何重建效果。与 People-Snapshot[131]相比，本方法可以生成更加精细的几何形状，且能够处理穿着松散衣服的人体。

People-Snapshot[131]重建得到的模型的上衣紧贴人体。People-Snapshot 数据集中的一些场景是在户外环境中拍摄的，此类数据具有强烈的光照变化。本方法在这些数据上仍然得到了照片级逼真的渲染结果，这表明本方法可以处理复杂的光照条件。

三维重建比较：图 3-7 呈现了本方法和 People-Snapshot[131]三维重建的定性结果。本方法比 People-Snapshot[131]恢复了更多的几何细节。例如，头发的形状与 RGB 观察高度一致。图中最后一列的结果表明本方法可以重建出穿着宽松衣服的人体几何模型，而 People-Snapshot[131]难以恢复此类数据的正确形状。

表 3-4 ZJU-MoCap 数据集视频序列 “Twirl” 上在不同视角数目训练的模型的量化比较。本章选择 6 个视角进行该消融实验，其余视角用于测试。

	1 个视角	2 个视角	4 个视角	6 个视角
PSNR	25.08	25.49	30.54	**32.73**
SSIM	0.912	0.928	0.969	**0.979**

表 3-5 视频序列 “Twirl” 上使用不同视频帧数训练的模型的量化比较。本章选择 1 帧、60 帧、300 帧、600 帧、1200 帧进行该消融实验。第 1 帧用于测试在训练人体姿态上的新视角合成质量。

	1 帧	60 帧	300 帧	600 帧	1200 帧
PSNR	25.64	30.14	**30.66**	30.59	29.97
SSIM	0.940	0.970	**0.971**	0.970	0.970

3.3.3 ZJU-MoCap 数据集上的消融实验

本章在 ZJU-MoCap 数据集的视频序列 “Twirl” 上进行了消融实验，探索了各个模块设计和输入数据对模型性能的影响，包括逐帧隐变量、视角数量、输入视频长度、隐变量扩散方法、目标函数。

逐帧隐变量的影响：为了验证第 3.2.4 节中提出的逐帧隐变量 $\{\boldsymbol{\ell}_t\}_{t=1}^{N_t}$ 的有效性，本章额外训练了一个没有逐帧隐变量的模型。该模型在新视角合成上给出了 30.03 PSNR，低于完整模型的 30.56 PSNR。这个实验结果表明逐帧隐变量提升了 0.53 PSNR。

输入视角数量的影响：表 6-2 比较了使用不同数量的相机视角进行训练的模型。结果表明，增加训练视角的数量提高了模型在新视角合成上的性能。值得注意的是，本模型即使只在单视角上训练，性能仍然优于使用四个视图训练的 Neural Volumes[81]。Neural Volumes 在该消融实验的测试视角上只给出了 23.12 PSNR 和 0.875 SSIM。

视频长度的影响：为了探索视频长度对模型性能的影响，本实验分别使用 1 帧、60 帧、300 帧、600 帧、1200 帧来训练本章提出的模型。表 3-5 呈现了定量结果，表明在视频上进行训练可以提高视图合成性能。但是在过多的视频帧上训练可能会降低模型在

表 3-6 ZJU-MoCap 数据集视频序列 “Twirl” 上使用不同的扩散方法的模型的量化比较。训练迭代时间是指每次迭代训练所花费的时间。

	PSNR	SSIM	训练迭代时间
PointNet	18.23	0.797	**0.1045 秒**
PointNet++	26.05	0.931	0.8555 秒
三维卷积神经网络	**30.54**	**0.969**	0.1748 秒

训练帧上的新视角合成性能，因为网络难以拟合非常长的视频。

隐变量扩散方法的影响：第 3.2.3 节提出使用三维卷积神经网络将 SMPL 模型表面上的隐变量扩散到表面附近的三维空间中。为了验证其有效性，本方法将其与其他两种融合方式进行比较：（1）PointNet[242]。对于任何点，此扩散方式找到所有距离该点在一定半径范围内的 SMPL 顶点（最多 K 个顶点），并对这些顶点的隐变量应用 PointNet[242]以获取点特征。（2）PointNet++[144]。此方式首先使用 PointNet++ 网络对结构化隐变量提取分层的特征向量。然后，对于任何三维点，此方式使用多尺度分组来索引不同尺度的特征向量，并将这些向量连接起来形成一个多尺度的特征向量。

表 3-6 比较了三种隐变量扩散方法在图像合成质量和训练期间的运行时间方面的表现。在 PSNR 和 SSIM 指标方面，三维卷积网络明显优于其他两种隐变量扩散方法。而在训练时间方面，三维卷积网络比 PointNet++ 快得多。

图像感知损失函数的影响：如第 3.2.5 节所述，本方法采用了 MSE 损失函数来训练网络模型。本消融实验探索了更复杂的损失函数对模型性能的影响，比如 NHR[25]中的图像感知损失。本实验同时使用图像感知损失函数和 MSE 损失函数监督训练了一个模型。实验结果显示，该模型在 PSNR 和 SSIM 指标上的渲染性能与原始模型类似（PSNR：30.70 vs. 30.54，SSIM：0.970 vs. 0.969）。

3.4 总结与讨论

本章提出了了一种新颖的基于隐式神经表示的动态人体模型，用于从稀疏视角的视频中合成动态人体的自由视角视频。本模型定义了一组结构化隐变量，与 MLP 网

络结合用于编码人体的局部几何和纹理。该模型将这组隐变量锚定在一个参数化人体模型的顶点上，通过参数化人体模型来变换隐变量的空间位置，从而能建模动态人体。这相当于建立了一个隐变量模型，从同一组隐变量中生成不同视频帧的数字人体，因此可以有效地整合跨视频帧的观测结果。本章在输入的视频上通过可微分体积渲染联合优化了上述提出的动态人体模型。为了评估该方法，本章创建了一个多视角数据集，使用相机阵列拍摄复杂运动中的动态人体。在新收集的数据集和 People-Snapshot 数据集上的实验结果表明，本章提出的方法相比于之前的方法展示了卓越的新视角合成质量。值得一提的是，本章工作启发了诸多后续工作，比如 Neural Human Performer[67]、GP-NeRF[243]、Relighting4D[91]。除此之外，本章创建的 ZJU-MoCap 数据集被大量后续工作[90,217-218,244-246] 使用，大大促进了数字人领域的发展。

第 4 章　基于骨骼蒙皮驱动的人体神经辐射场表示

4.1　引言

本文第 3 章提出的隐式神经表示实现了从稀疏视角视频重建动态人体的自由视角视频，大大降低了人体建模所需的硬件成本。然而，第 3 章未曾考虑数字人驱动性的问题，其提出的神经人体模型在新人体姿态下的渲染性能表现不佳。可驱动的数字人是自由视角视频、远程会议、游戏制作等实际应用的关键技术。为此，本章设计了一种基于骨骼蒙皮驱动算法的动态数字人表示，致力于合成高质量的新人体姿态下的图片。

制作可驱动数字人的传统管线往往是比较昂贵和耗时的。这由两个原因导致：首先，如在第 3 章所描述的，为了建模高质量的数字人体，传统方法通常依赖于复杂且昂贵的硬件设备，例如稠密的摄像机阵列[1,96] 或深度传感器[7,9]。其次，之前的方法大多基于骨骼蒙皮驱动算法[21] 构建可驱动的人体模型。这需要熟练的三维艺术家手动创建适用于特定人体的骨骼模型，并仔细设计三维网格模型每个顶点的蒙皮权重[21]来实现逼真自然的驱动效果。这个过程需要大量的人力和时间。本章的目标是降低建模可驱动数字人的成本，以实现大规模的人体数字化。具体而言，本章致力于从一段稀疏多视角视频中重建可驱动的人体模型。图 4-1 给出了一个示例。为了完成这个任务，本章探索了如何表示可驱动的数字人以及如何从 RGB 视频中学习这种人体表示。

近年来，基于神经辐射场 NeRF 的研究工作[15,72] 在静态三维场景的重建与渲染上取得了非常好的效果。为了将神经辐射场扩展到能处理非刚性动态场景，[247-248]将动态场景分解为标准坐标系下的神经辐射场和每一时刻的空间变形场。这个变形场将每个时刻

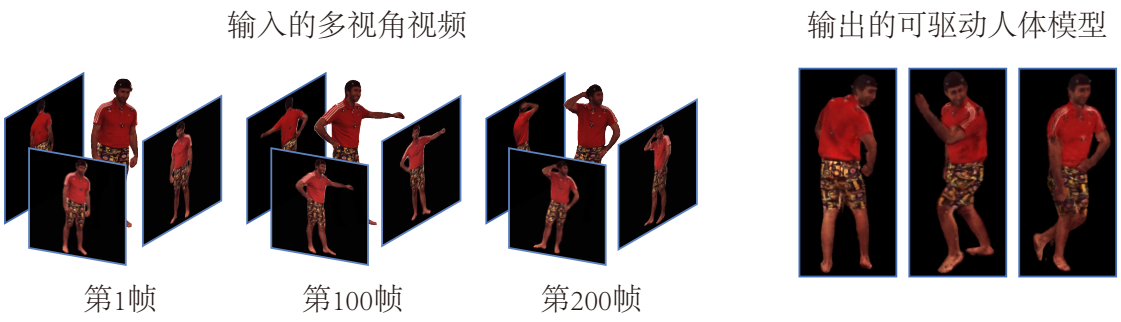


图 4-1 本章提出了一种基于骨骼蒙皮驱动模型的动态数字人表示，可以从多视角视频中自动重建可驱动的数字人体模型，从而用于合成新人体姿态下的图片。

世界坐标下的任意三维点变换到标准坐标系。之前的研究工作使用平移矢量场[248]或者 SE(3) 场[247]表示空间变形场。尽管这些研究工作可以处理一些动态场景，但由于两个原因，此类工作不适合表示可驱动的数字人模型。首先，这些工作基于可微分渲染从图片中优化神经辐射场和空间变形场的参数。但是在没有运动先验的情况下，同时优化神经辐射场与空间变形场面临较大的解空间，是一个欠定问题[248-249]，导致难以恢复正确的神经辐射场和变形场。其次，这些工作将时间作为输入以预测每一时刻的变形场，因此无法通过输入人体姿态来显式地驱动数字人。

为了解决上述提到的问题，本章提出了一种新颖的变形场表示方法。本方法首先使用人体姿态检测算法[22] 从稀疏视角图片中获得人体姿态参数。然后，本方法定义了一个基于隐身神经表示的神经蒙皮权重场，使用一个全连接神经网络预测任意一个三维点的蒙皮权重向量。最后，本方法基于骨骼蒙皮驱动算法将人体姿态参数与蒙皮权重场相结合，从而得到每个三维点从世界坐标系到标准坐标系的变换矩阵。本章提出的变形场表示有两个优点。首先，由于人体姿态参数易于跟踪[22]，因此不需要与神经蒙皮权重场共同优化，因此对变形场的学习提供了有效的正则化。而且基于预测得到的人体姿态参数，本方法可以利用参数化人体模型[104] 得到每一个时刻下的初始蒙皮权重场，进一步提升优化的稳定性。其次，本方法可以得到任意人体姿态下的变形场。具体的实现方式为，本方法利用了蒙皮权重场的循环一致性的特点，也即标准坐标系和世界坐标系相对应的两个三维点具有相同的蒙皮权重向量。基于此特点，本方法在标准坐标系下学习一个额外的神经蒙皮权重场，当给定一个新的人体姿态时，再根据循环一致性学习目标人体姿态下的神经蒙皮权重场，从而得到目标姿态下的变形场。

综上所述，本章的主要贡献可以概括为以下几点：首先，本章基于骨骼蒙皮驱动算法提出了一种新颖的数字人体表示，通过将神经辐射场与三维人体姿态结合来产生关联标准空间和观测空间的变形场，可以从稀疏视角视频中恢复高质量的数字人体模型，并支持使用三维人体姿态显式地驱动人体模型。其次，本章利用三维人体姿态与参数化人体模型得到训练时各个人体姿态下的初始蒙皮权重场，用于约束神经蒙皮权重场的优化，并通过循环一致性获得标准空间坐标系下的神经蒙皮权重场。最后，本章在 Human3.6M[23] 和 ZJU-MoCap[250] 数据集上进行了充分的实验以证明本章提出的方法的有效性。实验结果表明，该方法在新视图合成和新姿态合成方面表现出最先进的性能。此外，该方法可以从视频中重建出可驱动的三维人体网格模型。

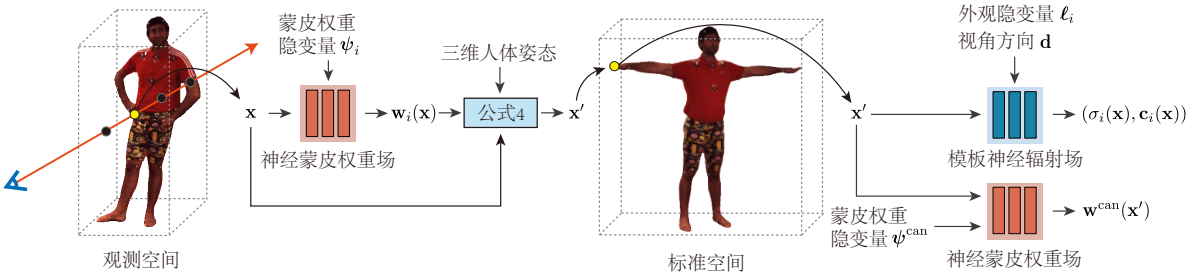


图 4-2 方法概述。给定在第 i 帧观测空间中的一个查询点 $\mathbf{x}$，本方法使用一个神经蒙皮权重场来推断其蒙皮权重 $\mathbf{w}_i(\mathbf{x})$，该 MLP 网络的输入是第 i 帧的隐变量 ψ_i。基于蒙皮权重和三维人体姿态，本方法使用公式 (4-4) 来计算出查询点在标准空间中的对应点 $\mathbf{x}'$。本方法将标准空间中的点 $\mathbf{x}'$、观测空间中的视角 $\mathbf{d}$ 和第 i 帧的外观隐变量 ℓ_i 作为输入，将其输入到模板神经辐射场中，从而预测体素密度和颜色。为了实现可驱动的人体模型，本方法还在标准空间中学习了一个神经蒙皮权重场 $\mathbf{w}^{\text{can}}(\mathbf{x}')$。

4.2　方法

4.2.1　方法概述

给定一个动态人体的多视角视频，本方法的目标是重建一个可驱动的数字人体模型，从而可用于合成数字人体在新的动作序列下的自由视点视频。本方法假定了输入视频的相机参数已知，并且视频的每一帧已经给定了三维人体姿态参数。这些人体姿态可以使用人体姿态估计系统[22-23]来获得。对于每一张图像，本方法使用人体分割算法[232]来提取前景人体掩模，用于去除输入图片的背景。

图 4-2 展示了本章提出的方法流程图。本章将动态数字人体分解成由神经辐射场表示的标准人体模型（第 4.2.2 节）和蒙皮权重场（第 4.2.3 节），用于建立标准空间和观测空间之间的对应关系。然后本章讨论了如何在多视角视频上学习上述提出的人体模型（第 5.2.4 节）。基于蒙皮权重场，本方法实现了可驱动的数字人体模型（第 4.2.5 节）。

4.2.2　基于神经辐射场的动态场景表示

神经辐射场将静态三维场景表示为一个 MLP 网络。对于任何空间三维点，神经辐射场将空间位置 $\mathbf{x}$ 和视角方向 $\mathbf{d}$ 作为 MLP 网络的输入，并输出体素密度 σ 和颜色 $\mathbf{c}$。

受到之前研究工作[247-248]的启发，本方法引入了基于隐式神经表示的变形场来扩展神经辐射场的能力，从而可以表示动态人体。具体而言，对于每个视频帧 $i \in \{1, ..., N\}$，本方法定义了一个变形场 T_i，用于将观测空间的三维点变换到标准空间。给定标准空间

下的体素密度模型 F_σ，第 i 帧的体素密度模型被定义为：

$$(\sigma_i(\mathbf{x}), \mathbf{z}_i(\mathbf{x})) = F_\sigma(\gamma_\mathbf{x}(T_i(\mathbf{x}))), \tag{4-1}$$

其中 $\mathbf{z}_i(\mathbf{x})$ 是一个高维特征向量，而 $\gamma_\mathbf{x}$ 是对于空间三维点的位置编码函数[15]。

当预测三维点颜色时，本方法定义了逐帧的可学习特征向量 $\boldsymbol{\ell}_i$ 来编码第 i 帧的人体外观状态。给定标准空间的颜色模型 $F_\mathbf{c}$，第 i 帧的颜色模型可以定义如下：

$$\mathbf{c}_i(\mathbf{x}) = F_\mathbf{c}(\mathbf{z}_i(\mathbf{x}), \gamma_\mathbf{d}(\mathbf{d}), \boldsymbol{\ell}_i), \tag{4-2}$$

其中 $\gamma_\mathbf{d}$ 是用于视角方向的位置编码函数。

之前的研究工作已经提出了一些变形场的隐式神经表示，如平移向量场[248-249] 和 SE(3) 场[247]。然而，正如之前的方法[247,249] 所讨论的，基于输入视频同时优化神经辐射场和变形场是一个欠定的问题，容易陷入局部最优解。为了克服这个问题，这些方法[247,249]提出了许多正则化的技术来提升训练的稳定性，但这使得训练过程变得较为复杂和受限。除此之外，这些方法的变形场表示无法被输入的三维人体姿态显式地驱动，因此无法生成新的运动序列下的神经辐射场。

4.2.3 神经蒙皮权重场

考虑到本章的目标是建模可驱动的人体模型，因此利用人体先验来帮助变形场的建模是很自然的思路，有助于解决训练过程欠约束的问题。具体而言，本工作基于三维人体姿态和骨骼蒙皮驱动模型[21] 构建变形场。

人体骨架定义了 K 个骨骼，这些骨骼用于产生 K 个变换矩阵 $G_k \in SE(3)$。在骨骼蒙皮驱动算法[21]中，标准空间的三维点 $\mathbf{v}$ 被变换到观测空间的过程被定义为：

$$\mathbf{v}' = \left(\sum_{k=1}^{K} w(\mathbf{v})_k G_k \right) \mathbf{v}, \tag{4-3}$$

其中 $w(\mathbf{v})_k$ 是第 k 部分的蒙皮权重。同样的，对于一个观测空间中的三维点 $\mathbf{x}$，如果已知其对应的蒙皮权重，骨骼蒙皮驱动算法就能使用公式将其转换到标准空间：

$$\mathbf{x}' = \left(\sum_{k=1}^{K} w^o(\mathbf{x})_k G_k \right)^{-1} \mathbf{x}, \tag{4-4}$$

其中 $w^o(\mathbf{x})$ 是在观测空间中定义的蒙皮权重函数。为了获得蒙皮权重场，一个自然的想法是定义一个将三维坐标点映射到蒙皮权重的 MLP 网络，然后根据方程(4-1)，

(4-2)和(4-4)得到动态神经辐射场。然而，本工作发现联合优化标准坐标系下的神经辐射场和蒙皮权重场是欠定的，并且容易陷入局部最小值。

为了解决这个问题，本工作借助人体参数化模型[104,108-109,251]中获取人体先验，从而来正则化蒙皮权重场的训练过程。具体来说，对于任何三维点，本工作根据参数化人体模型分配初始的蒙皮权重，然后使用 MLP 网络预测残差权重向量，将两者相加得到最终的神经蒙皮权重场。在实际实现中，所有训练视频帧的残差蒙皮权重场使用单个 MLP 网络 $F_{\Delta\mathbf{w}}:(\mathbf{x},\boldsymbol{\psi}_i)\rightarrow\Delta\mathbf{w}_i$ 来实现，其中 $\boldsymbol{\psi}_i$ 是逐帧可学习的特征向量，$\Delta\mathbf{w}_i$ 是残差蒙皮权重向量 $\in\mathbb{R}^K$。第 i 帧的神经蒙皮权重场定义为:

$$\mathbf{w}_i(\mathbf{x})=\text{norm}(F_{\Delta\mathbf{w}}(\mathbf{x},\boldsymbol{\psi}_i)+\mathbf{w}^{\text{s}}(\mathbf{x},S_i)), \tag{4-5}$$

其中 $\mathbf{w}^{\text{s}}$ 是基于人体参数化模型 S_i 计算出的初始蒙皮权重。本工作定义 $\text{norm}(\mathbf{w})=\mathbf{w}/\sum w_i$。在不失一般性的情况下，本工作采用 SMPL[104]作为人体参数化模型。SMPL 模型参数可以通过将 SMPL 模型拟合到 3D 人体骨架[22]来获得。需要注意的是，这个想法也适用于其他人体参数化模型[108-109,251]。为了计算初始蒙皮权重 $\mathbf{w}^{\text{s}}$，本工作采用了之前研究工作[60,252]提出的策略。对于任何三维坐标点，本工作首先找到 SMPL 三维网格模型上最近的表面点，然后通过对相应网格面三个顶点的蒙皮权重进行重心插值（Barycentric interpolation），从而得到目标蒙皮权重。

为了通过人体姿态参数驱动标准坐标系下的神经辐射场，本工作在标准空间中额外定义了一个神经蒙皮权重场 $\mathbf{w}^{\text{can}}$。为此，本工作使用 T 姿势下的 SMPL 模型计算得到 SMPL 蒙皮权重场 $\mathbf{w}^{\text{s}}$，并且定义了一个可学习特征向量 $\boldsymbol{\psi}^{\text{can}}$，将其输入 $F_{\Delta\mathbf{w}}$ 预测残差蒙皮权重场。本方法利用了标准空间和观测空间的蒙皮权重场的一致性来训练标准空间下的神经蒙皮权重场 $\mathbf{w}^{\text{can}}$。具体训练过程将在第 5.2.4 节中描述。

相比于在观测空间与标准空间同时学习蒙皮权重场，另一种方式是只在标准空间学习蒙皮权重场。基于公式(4-3)，标准空间的蒙皮权重场定义了从标准空间到观测空间的匹配，这是一个前向的变形过程。然而，基于这个公式(4-3)以获取从观测空间到标准空间的匹配并不直接。为了实现这个目标，首先需要通过密集采样观测空间上的三维点并评估这些三维点的蒙皮权重，以此建立一组密集的观测空间到标准空间的匹配。然后，对于任意的观测空间的三维点，可以通过这组预先计算的密集匹配来获得其对应的标准空间的三维点。上述这个过程较为复杂而且计算量大。除此之外，由于采样点是离散的，

所以这个过程计算出的匹配往往较为粗糙。对比之下，通过在观测空间中定义蒙皮权重场，本工作可以根据公式(4-4)直接获得从观测空间到标准空间的匹配。

4.2.4 模型训练细节

给定动态神经辐射场 σ_i 和 $\mathbf{c}_i$，本工作使用体积渲染技术[15,253]来合成每个视频帧 i 在任意视角下的图像。本方法利用 SMPL 模型计算场景的边界框，以此设定体积渲染的近平面和远平面。通过最小化渲染像素颜色 $\tilde{\mathbf{C}}_i(r)$ 和观测像素颜色 $\mathbf{C}_i(r)$ 之间的差异，本工作同时优化了体素密度模型 F_σ、颜色模型 $F_\mathbf{c}$、残差蒙皮权重场 $F_{\Delta\mathbf{w}}$、可学习特征向量 $\boldsymbol{\ell}_i$ 和 $\boldsymbol{\psi}_i$ 的参数。渲染误差的公式被定义为：

$$L_{\text{rgb}} = \sum_{r\in\mathcal{R}} \|\tilde{\mathbf{C}}_i(\mathbf{r}) - \mathbf{C}_i(\mathbf{r})\|_2, \tag{4-6}$$

其中 $\mathcal{R}$ 是图片像素对应的相机射线集合。

为了训练在标准空间中的神经蒙皮权重场 $\mathbf{w}^{\text{can}}$，本工作额外使用了基于蒙皮权重一致性的损失函数。根据公式(4-3)和(4-4)可知，标准空间和观测空间的两个对应的三维点应该具有相同的蒙皮权重。对于第 i 帧的观察空间中的三维点 $\mathbf{x}$，本工作使用公式(4-4)将其映射为标准空间中的三维点 $T_i(\mathbf{x})$。蒙皮权重场之间的一致性损失函数被定义为：

$$L_{\text{nsf}} = \sum_{\mathbf{x}\in\mathcal{X}_i} \|\mathbf{w}_i(\mathbf{x}) - \mathbf{w}^{\text{can}}(T_i(\mathbf{x}))\|_1, \tag{4-7}$$

其中 $\mathcal{X}_i$ 是第 i 帧的观测空间中采样的三维点集合。在本章实验中，L_{rgb} 和 L_{nsf} 被设为 1。

4.2.5 人体模型驱动

为了合成数字人在新的人体姿态下的图片，本工作在相应的姿态空间下构造神经蒙皮权重场，用于产生变形场将姿态空间下的三维点变换到标准空间。给定一个新的人体姿态，本方法首先计算对应的 SMPL 人体模型，并根据 SMPL 参数 S^{new} 计算 SMPL 蒙皮权重场 $\mathbf{w}^{\text{s}}$。然后，新人体姿态的神经蒙皮权重场 $\mathbf{w}^{\text{new}}$ 被定义为：

$$\mathbf{w}^{\text{new}}(\mathbf{x}, \boldsymbol{\psi}^{\text{new}}) = \text{norm}(F_{\Delta\mathbf{w}}(\mathbf{x}, \boldsymbol{\psi}^{\text{new}}) + \mathbf{w}^{\text{s}}(\mathbf{x}, S^{\text{new}})), \tag{4-8}$$

其中 $\boldsymbol{\psi}^{\text{new}}$ 是新人体姿态的可学习特征向量。基于这个神经蒙皮权重场和公式(4-4)，本方法使用骨骼蒙皮驱动算法计算得到目标人体姿态下的空间变形场 T^{new}。可学习特征

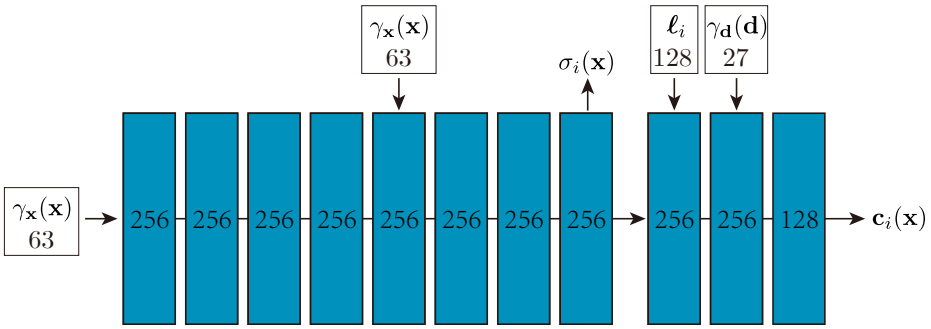


图 4-3 神经体素密度场和颜色场的网络结构。相比于 NeRF[15] 的网络，本方法引入了一个逐帧的隐编码 $\boldsymbol{\ell}_i$ 来记录第 i 帧中人体的外观状态。每个网络层中的数字表示其输入的维度。

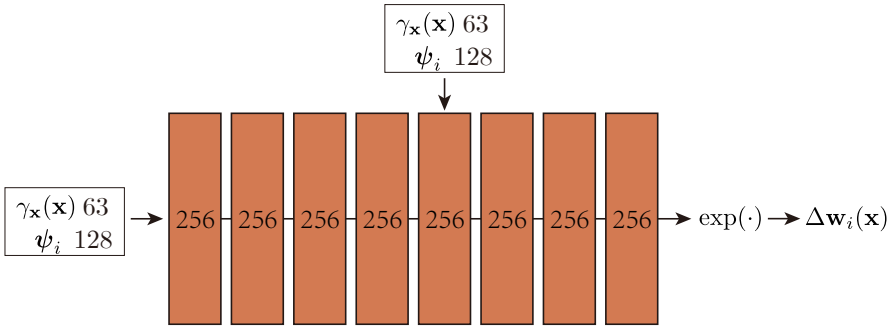


图 4-4 神经蒙皮权重场的网络结构。网络的输入是三维点的位置编码 $\gamma_{\mathbf{x}}(\mathbf{x})$ 和逐帧的隐编码 $\boldsymbol{\psi}_i$，输出是残差蒙皮权重 $\Delta\mathbf{w}_i(\mathbf{x})$。网络结构包含 8 个线性层，每层包含了一个 ReLU 激活函数。

向量 $\boldsymbol{\psi}^{\text{new}}$ 基于这个目标函数被优化：

$$L_{\text{new}} = \sum_{\mathbf{x}\in\mathcal{X}^{\text{new}}} \|\mathbf{w}^{\text{new}}(\mathbf{x}) - \mathbf{w}^{\text{can}}(T^{\text{new}}(\mathbf{x}))\|_1, \tag{4-9}$$

其中 $\mathcal{X}^{\text{new}}$ 是在新的人体姿态空间的三维点集合。请注意，在训练过程中，本工作固定了标准空间中神经蒙皮权重场 $\mathbf{w}^{\text{can}}$ 的参数。通过变形场 T^{new}，本方法使用方程(4-1)和(4-2)生成新的人体姿势下的神经辐射场。

4.3 实现细节

本章中的神经体素密度场 F_σ 和颜色场 $F_{\mathbf{c}}$ 遵从了 NeRF[15] 的实现。$F_{\Delta\mathbf{w}}$ 的网络结构与 F_σ 几乎相同，除了最后一层输出通道数为 24。具体网络结构细节见图 4-3 和 4-4。外观隐编码 $\boldsymbol{\ell}_i$ 和蒙皮权重场隐编码 $\boldsymbol{\psi}_i$ 的维度均为 128。在体积渲染过程中，本方法只使用了单阶段的三维点采样策略，沿着每条相机射线采样 64 个点。

训练：本章提出的方法采用了两阶段训练流程。首先，本方法在输入的视频上联合训练 F_σ、$F_{\mathbf{c}}$、$F_{\Delta\mathbf{w}}$、$\{\boldsymbol{\ell}_i\}$ 和 $\{\boldsymbol{\psi}_i\}$ 的参数。其次，本方法使用方程 (4-9) 在新的人体姿态下学习神经蒙皮权重场。本章实验使用 Adam 优化器[240]进行训练。学习率从 $5e^{-4}$ 开始，指数衰减到 $5e^{-5}$。训练在四块 2080 Ti GPU 上进行。对于 300 帧的三视角视频，第一阶段训练大约需要 20 万次迭代（大约 12 小时）。对于 200 个新的人体姿态，第二阶段训练大约需要 1 万次迭代（大约 30 分钟）。

4.4 实验分析

4.4.1 数据集和实验指标

Human3.6M 数据集[23]：该数据集使用 4 个同步摄像机拍摄多视角视频，并使用基于标记的人体运动捕捉系统收集人体姿态。数据集拍摄的动态人体执行了多样且复杂的人体动作。本工作选择其中的代表性动作，将视频分成训练和测试帧，并在编号为 S1，S5，S6，S7，S8，S9 和 S11 的视频数据上进行实验，使用 3 个视角进行模型训练，并选择剩余摄像机进行测试。

ZJU-MoCap 数据集[250]：该数据集由第 3 章中的工作采集，其使用了 21 个相机拍摄多视角视频，并使用无标记运动捕捉系统[22] 采集人体姿态。本章实验选择了四个具有代表性的动作序列进行实验，分别是 “旋转”，“太极”，“热身”、“拳击 1”。本实验遵循了第 3 章中的实验设定。

实验指标：本工作使用两个典型的实验指标[15]评估本章提出的模型在图像合成上的表现：峰值信噪比（PSNR）和结构相似性指数（SSIM）。

4.4.2 图像合成的实验结果

基线方法：本章将提出的方法与基于 SMPL 模型的图像合成方法[25,159,250]进行比较。（1）Neural Textures[159] 使用可学习的特征图和二维卷积神经网络来将粗糙的三维网格模型转化为目标图像。由于 Neural Textures[159]未开源，本工作重新实现了 Neural Textures 并将 SMPL 网格作为输入网格。（2）NHR[25] 从输入点云中提取 3D 特征并将其渲染成二维特征图，然后使用二维卷积神经网络转换成图像。由于从稀疏相机视图中获取密集点云很困难，本工作将 SMPL 顶点作为 NHR 的输入点云。

表 4-1 Human3.6M 数据集上的新视角合成结果。PSNR 和 SSIM 的指标数值越高越好。“NT” 表示 Neural Textures[159]。

	PSNR ↑			SSIM ↑		
	NT[159]	NHR[25]	本方法	NT[159]	NHR[25]	本方法
S1	20.98	21.08	**22.05**	0.860	0.872	**0.888**
S5	19.87	20.64	**23.27**	0.855	0.872	**0.892**
S6	20.18	20.40	**21.13**	0.816	0.830	**0.854**
S7	20.47	20.29	**22.50**	0.856	0.868	**0.890**
S8	16.77	19.13	**22.75**	0.837	0.871	**0.898**
S9	22.96	23.04	**24.72**	0.873	0.879	**0.908**
S11	21.71	21.91	**24.55**	0.859	0.871	**0.902**
平均	20.42	20.93	**23.00**	0.851	0.866	**0.890**

表 4-2 Human3.6M 数据集上的新人体姿态合成结果。“NT” 表示 Neural Textures[159]。

	PSNR ↑			SSIM ↑		
	NT[159]	NHR[25]	本方法	NT[159]	NHR[25]	本方法
S1	20.09	20.48	**21.37**	0.837	0.853	**0.868**
S5	20.03	20.72	**22.29**	0.843	0.860	**0.875**
S6	20.42	20.47	**22.59**	0.844	0.856	**0.884**
S7	20.03	19.66	**22.22**	0.838	0.852	**0.878**
S8	16.69	18.83	**21.78**	0.824	0.855	**0.882**
S9	22.20	22.18	**23.72**	0.851	0.860	**0.886**
S11	21.72	22.12	**23.91**	0.854	0.867	**0.889**
平均	20.17	20.64	**22.55**	0.842	0.858	**0.880**

图 4-5 Human3.6M 数据集上的新视角合成的量化结果。之前的方法[25,159] 在渲染的过程中难以控制视角，并且倾向于过拟合训练视角。与之相比，本工作能够准确地渲染目标视角。

新视角合成比较：为了进行各个方法在新视角合成性能的比较，本章在训练视频帧的新视角上合成图片。表 4-1 呈现了本章提出的方法与基线方法[25,159]的量化比较。具体来说，本章提出的模型在 PSNR 指标中的表现优于其他方法[25,159]至少 2.07，在 SSIM 度量中的表现优于其他方法[25,159]至少 0.024。

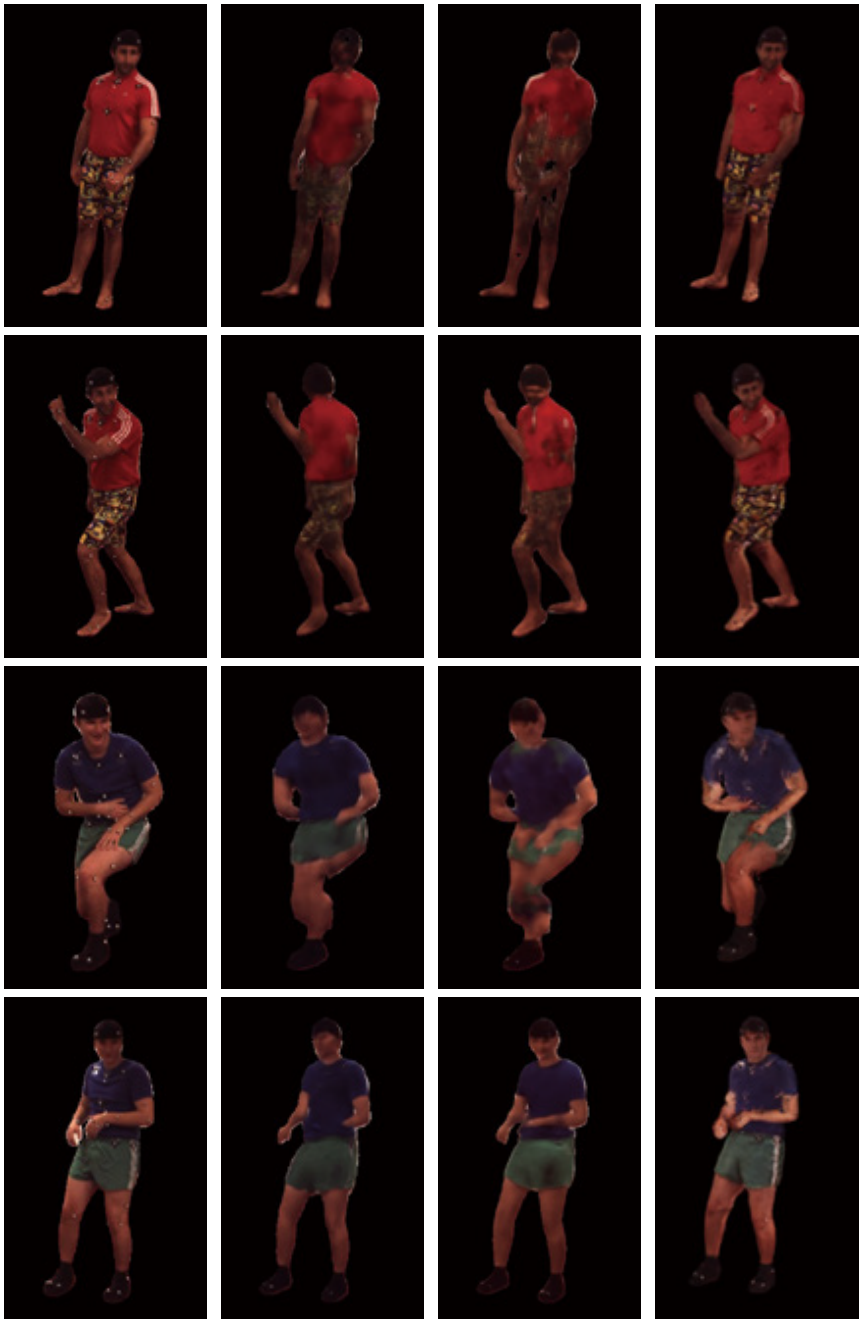


图 4-6 Human3.6M 数据集上的新人体姿态合成的定性结果。对于复杂的人体姿态，之前的方法[25,159,250]往往会生成变形的渲染结果。与之相比，本工作在复杂人体姿态的渲染结果更好。

图 4-5 展示了本文的方法和基线方法[25,159]的定性比较。基线方法[25,159]都难以控制渲染视角。从渲染结果可以看出这两个方法倾向于合成训练视角的内容。如图 4-5 中第二个人所示，Neural Textures 和 NHR 合成了了训练时看到的人体背面。相比之下，由于预测了三维人体表示，本方法能够准确地控制视角。

图 4-7 ZJU-MoCap 数据集上的新人体姿态合成的定性比较。

表 4-3 ZJU-MoCap 数据集上训练人体姿态和新人体姿态的新视角合成的量化比较。“NB” 是第 3 章中提出的数字人体模型。

	PSNR ↑				SSIM ↑			
	NT[159]	NHR[25]	NB	本方法	NT[159]	NHR[25]	NB	本方法
训练姿态	22.61	23.25	**28.90**	27.10	0.899	0.905	**0.967**	0.949
新姿态	21.55	21.88	23.06	**23.16**	0.860	0.863	0.879	**0.893**

新人体姿态合成比较：为了进行新人体姿态合成的比较，本章使用各个方法在测试视频帧的测试视角上合成图片以进行对比。表 4-2 将本方法与基线方法[25,159]在 PSNR 和 SSIM 指标方面进行了比较。对于这两个指标，本方法都给出了最佳性能。图 4-6 展示了定性比较。对于复杂的人体姿势，基线方法[25,159] 的渲染结果较为模糊。相比之下，本章提出的模型合成的图像质量更好。这些结果表明，本模型在图像生成过程中可以更好地控制渲染视角，优于基于二维卷积神经网络的方法。表 4-3 和图 4-7 显示，本章提出的模型在 ZJU-MoCap 数据集上的渲染质量也超过了第 3 章中提出的模型。

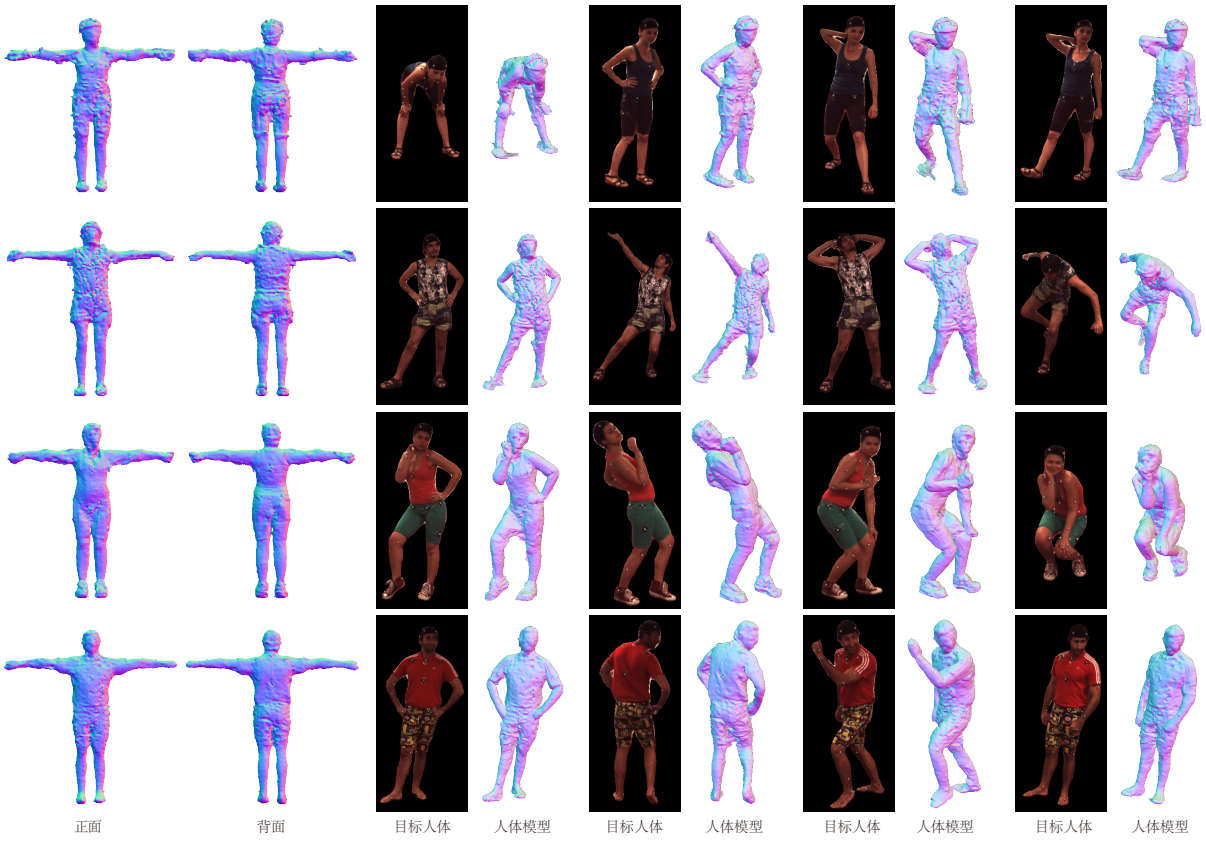


图 4-8 标准空间和观测空间下的人体三维模型。前两列呈现了标准空间下的人体三维模型，本方法通过骨骼蒙皮驱动算法将该模型变形到目标人体姿态空间。

4.4.3 三维重建的实验结果

除了可以合成新人体姿态的图像，本章提出的方法还可以生成新人体姿态下的三维重建结果。具体而言，本工作首先将标准空间下的人体三维场景离散化为一个三维体素网格，每个体素的大小为 $5mm \times 5mm \times 5mm$。然后本方法对所有体素进行体素密度的评估，使用 Marching Cubes 算法[241]提取人体三维多边形网格。为了驱动标准空间下的人体模型，本方法从神经蒙皮权重场 w^{can} 中推断出网格顶点的蒙皮权重。最后，给定一个新的人体姿态，本方法使用公式 (4-3) 对每个顶点进行变换，从而得到在目标人体姿态下的网格模型。图 4-8 呈现了标准空间和各个姿态下的人体三维模型。

4.4.4 消融实验

本章在 Human3.6M 数据集[23]中的一个视频序列（S9）上进行了消融实验，探究了各个设计对模型在新人体姿态合成上的性能的影响。首先，为了分析学习 $F_{\Delta \mathbf{w}}$ 的好处，

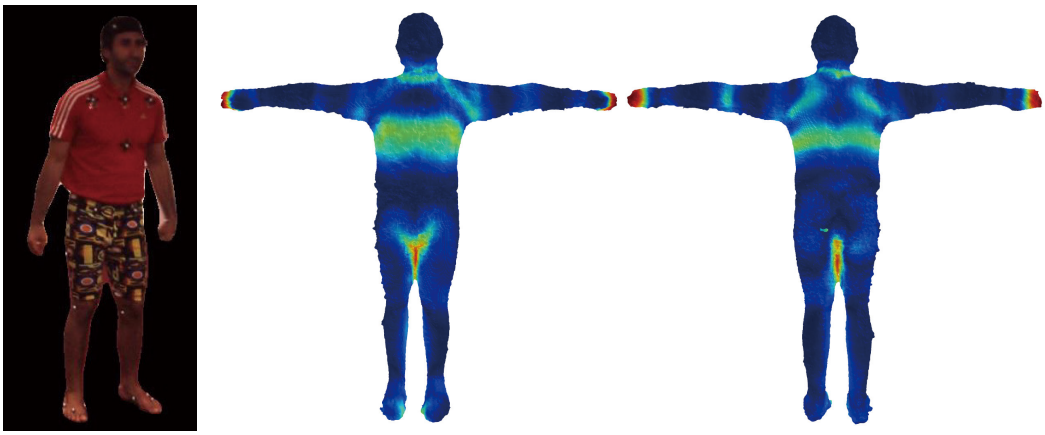


图 4-9 视频序列 “S9” 上优化得到的参差向量场 $F_{\Delta \mathbf{w}}$ 的可视化。颜色越红，残差越大。

表 4-4 视频序列 “S9” 上的神经蒙皮权重场和 SMPL 蒙皮权重场的新人体姿态合成结果。

	PSNR	SSIM
神经蒙皮权重场	**23.72**	**0.886**
SMPL 蒙皮权重场	21.65	0.850

本章将神经蒙皮权重场与 SMPL 蒙皮权重场进行比较。然后，为了探索输入的人体姿态精度对模型性能的影响，本章使用无标记人体捕捉方法[22,134] 估计了 SMPL 模型参数，并在这些参数上训练模型。最后，本章探索提出的方法在不同数量的视频帧和相机视角下的性能。表格 4-4、4-5、4-6 呈现了这些实验的结果。

神经蒙皮权重场的影响：表 4-4 呈现了定量比较，表明神经蒙皮权重场比 SMPL 蒙皮权重场表现更佳。图 4-9 更好地展示了在 SMPL 蒙皮权重场上的改进。本消融实验在标准空间中三维几何模型上可视化了残差向量场 $F_{\Delta \mathbf{w}}$。残差越大的区域颜色越红。实验结果表明，残差较大的区域主要位于脖子、手、胸部、裤子等区域。这些区域是 SMPL 模型的蒙皮权重难以描述的人体特定的几何细节。

输入人体姿态准确性的影响：表 4-5 比较了使用基于标记和无标记系统的人体姿态训练的模型。定性比较结果在图 4-10 中呈现。实验结果表明，更精确的人体姿态可以产生更高质量的渲染效果。值得一提的是，本章提出的模型在无标记系统采集的人体姿态上进行训练也能产生较好的图片合成效果。

图 4-10 视频序列 “S9” 上使用基于标记和无标记系统的人体姿态训练的模型的定量比较结果。

表 4-5 视频序列 “S9” 上的新人体姿态合成结果。两个模型分别使用基于标记检测系统和基于无标记检测系统获得的三维人体姿态进行训练。

	PSNR	SSIM
基于标记检测系统获得的三维人体姿态	**23.72**	**0.886**
基于无标记检测系统获得的三维人体姿态	22.27	0.858

表 4-6 视频序列 “S9” 上使用不同长度的视频进行训练的模型的新人体姿态合成结果。

	1 帧	100 帧	200 帧	800 帧
PSNR	20.29	23.40	**23.69**	23.16
SSIM	0.849	0.881	**0.883**	0.875

训练视频长度的影响：为了探究训练视频长度对模型性能的影响，本消融实验在同一个动作序列上分别使用 1 帧、100 帧、200 帧、800 帧的视频进行训练，并在剩余的测试帧上比较模型在新人体姿态合成方面的性能。表 4-6 列出了使用不同数量视频帧进行训练的模型的定量结果。结果表明，在视频上训练模型有助于恢复正确的三维人体表示，但网络似乎在拟合非常长的视频时有困难。根据实验经验，本工作发现在 150 到 300 帧的视频上训练能得到较好的结果。图 4-11 呈现了定性比较。

图 4-11 视频序列 “S9” 上使用不同数量视频帧进行训练的模型的定量比较结果。

表 4-7 视频序列 “S9” 上使用不同数量的视角进行训练的模型的新人体姿态合成结果。

	1 个视角	2 个视角	3 个视角
PSNR	23.81	**24.16**	23.72
SSIM	0.877	0.880	**0.886**

训练视角数量的影响：为了进行比较，本消融实验选择一个视角进行测试，并分别使用与该视角最接近的 1、2、3 个视角进行训练。表 4-7 比较了使用不同数量视角进行训练的模型的性能。实验结果表明，三个模型的定量性能相似。图 4-12 进一步比较了三个模型。结果表明，使用 3 个视角进行训练的模型渲染的细节更多。值得一提的是，使用单个视角进行训练的模型也能产生较好的渲染效果。

4.4.5 模型渲染速度

本章在 Human3.6M 数据集上测试渲染速度。对于 512×512 的图像，本章提出的模型在配备 Intel i7 3.7GHz CPU 和 GTX 1080 Ti GPU 的台式电脑上渲染该分辨率的图像需要 1.09s 的时间。具体来说，本模型预测颜色场和体素密度场需要 0.39s，预测蒙皮权重场需要 0.63s，进行体积渲染需要 0.07s。由于沿着光线采样的点的数量只有 64，并且人体的三维场景范围较小，因此本方法的渲染速度相对较快。

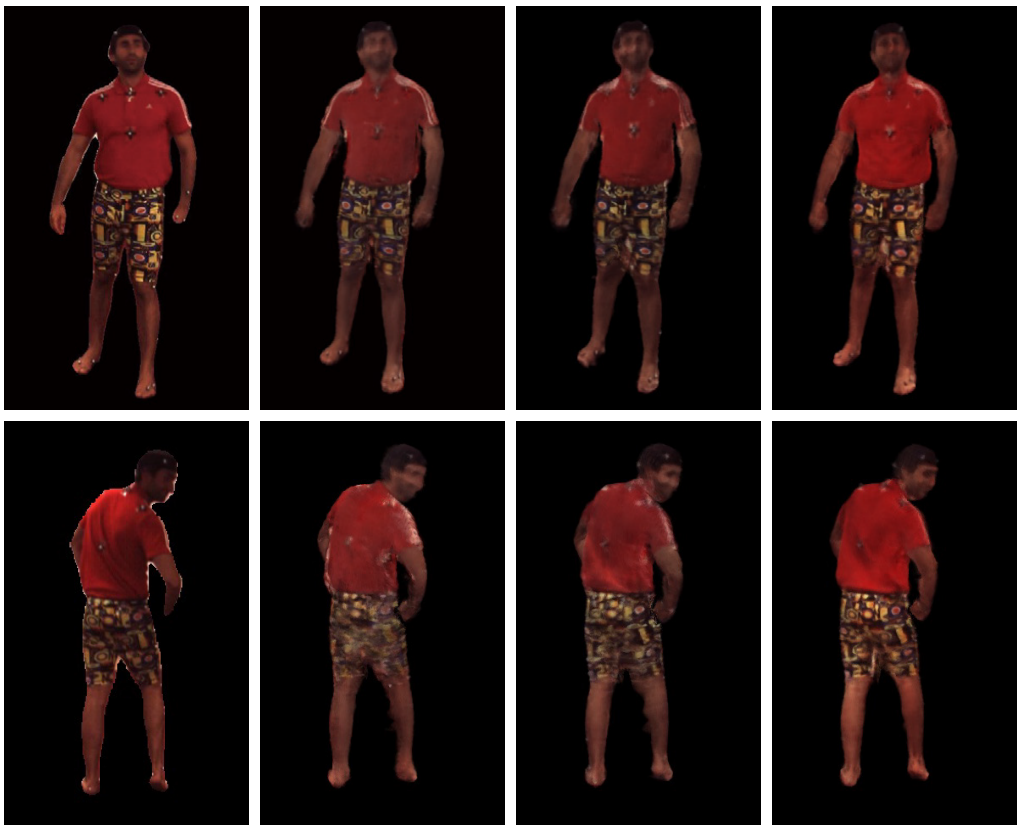


图 4-12 视频序列 “S9” 上使用不同数量视角进行训练的模型的定量比较结果。

4.5 总结与讨论

本章提出了一种新的动态人体表征方法，用于从多视角视频中建模可驱动人体模型。本方法通过神经变形场增强神经辐射场，使其可以建模可驱动的人体模型。具体而言，本方法基于骨骼蒙皮驱动模型定义了神经蒙皮权重场，与三维人体姿态相结合构建变形场，将观测空间的三维点转换到标准空间。可驱动的神经辐射场通过体积渲染在多视角视频上进行学习。本方法同时利用观测空间和标准空间的蒙皮权重场之间的一致性约束神经蒙皮权重场的优化过程。训练完成后，本方法可以合成新人体姿态序列下的自由视角视频。在 Human3.6M 数据集上的实验证明，本章所提出的模型在图像合成方面相比之前的方法达到了更好的渲染质量。除此之外，本工作启发了诸多的研究工作，比如 HumanNeRF[90]、ARAH[213]、NeuMan[254]。

第 5 章　基于符号距离场的人体几何表示

5.1　引言

本文的第 4 章提出了可驱动的数字人体模型，可以合成新人体姿态下的图片。然而，该方法采用神经辐射场建模人体几何，导致人体几何表面比较粗糙，和真实的人体几何相差较远。具有高质量几何的数字人在影视制作、虚拟试衣、虚实交互等应用中非常重要。比如进行虚拟试衣时，数字人的几何模型可以用于仿真衣物置于人体时的垂坠感以及衣物随人体的运动，从而让人们体验到逼真的试衣效果。

传统的人体建模算法[1,7-9] 可以重建出高精度的人体几何模型。这些研究工作将人体几何表示为三维网格，并将外观存储在网格模型对应的二维纹理图中。例如，基于稠密的相机阵列，Guo 等人[1] 使用多视图立体匹配方法[96-97] 重建三维网格模型，并基于球面梯度光照（Spherical gradient illumination）[255] 获得纹理图。尽管这个管线实现了高质量的重建和渲染结果，但该工作需要稠密相机阵列来进行多视图立体匹配。

近期的一些研究工作[85-86,248,250] 通过将时序信息整合到一个共享的人体模型中，从而实现基于稀疏的多视角视频中重建数字人体模型。第 4 章提出的 Animatable NeRF[86] 是其中的一个代表性工作，将人体模型表示为标准坐标系下的神经辐射场[15]。为了从 RGB 视频中学习这个标准坐标系下的人体模型，该工作定义了空间变形场，将世界坐标系下的三维点变换到标准坐标系，从而实现标准人体模型在图片空间的渲染。类似于神经辐射场，该工作通过最小化渲染图像和观察图像之间的差异来优化模型参数。尽管 Animatable NeRF 可以从视频中恢复支持高质量渲染的数字人模型，但该工作存在两个

图 5-1 本章提出了一种基于符号距离场的动态人体几何建模方法，该方法可以从单目视频中恢复高质量的三维人体几何，优于第 4 章中提出的建模方法。

缺陷。首先，该工作使用基于隐式神经表示的体素密度场建模人体几何。因为体素密度场缺乏表面的约束，导致人体的几何表面往往比较粗糙，和真实的人体几何相差较远。图 5-1 展示了一个例子。其次，为了在特定的人体姿态下渲染图像，该工作需要优化目标人体姿态下的神经蒙皮权重场。这使得数字人体的驱动过程较为不便。

本章提出了一种基于符号距离场（Signed distance field）的动态人体几何模型，称为 Animatable SDF。具体而言，本方法使用符号距离场表示标准坐标系下的人体几何。与体素密度场相比，符号距离场在零水平集上有明确定义的表面。并且符号距离场需要满足程函方程（Eikonal equation）的约束，从而对几何模型的学习过程有直接的正则化。然而，从 RGB 视频中优化标准坐标系下的符号距离场是一个未被解决的问题。渲染符号距离场的一个经典方式是球体追踪[16,24,150]。该渲染方法首先对于目标场景设定一个球面，计算球面上一个三维点的符号距离，然后沿着一条相机射线前进相应的距离，随后又计算所在位置的符号距离并再次前进，循环这个过程直到找到场景的表面点。虽然球体追踪在静态场景上取得了很好的效果，该渲染方法难以处理变形后的符号距离场[256]。这是因为世界坐标系中三维点的符号距离需要从标准坐标系获取，而标准坐标系和世界坐标系之间存在复杂的人体运动，导致了世界坐标系下的符号距离不一定是正确的[256]，也就导致球体追踪难以找到正确的表面点。为了解决这个问题，本章利用基于符号距离场的体积渲染技术[17-18] 来合成图片。实验证明这种方案可以有效地从视频中学习标准坐标系下的符号距离场，从而生成高质量的三维几何。

第 4 章提出的模型的另一个问题是驱动人体模型时需要优化目标人体姿态的神经蒙皮权重场，然后才能基于骨骼蒙皮驱动算法[21] 计算从世界坐标系到标准坐标系的变形场。解决这个问题的一个方法是直接从参数化人体模型任意三维点的蒙皮权重。具体过程是先获得世界坐标系下的参数化人体模型[104,108-109]，然后在模型表面找到离目标三维点最近的表面点，最后将该表面点的蒙皮权重作为目标三维点的蒙皮权重。然而，参数化人体模型通常仅描述未着衣物的人体，因此其蒙皮权重很可能无法准确地建模人体衣物的运动。为了处理衣物的运动，本章受到[85,257] 的启发，将人体运动分解为铰链式变形和非刚性变形。其中铰链式变形使用基于参数化人体模型的骨骼蒙皮驱动算法进行建模，而非刚性变形由一个基于隐式神经表示的位移场表示。

综上所述，本章贡献如下：首先，本章提出了一个基于符号距离场的动态人体几何表示，并通过可微分的体积渲染技术从稀疏视角的 RGB 视频中重建得到高质量的人体

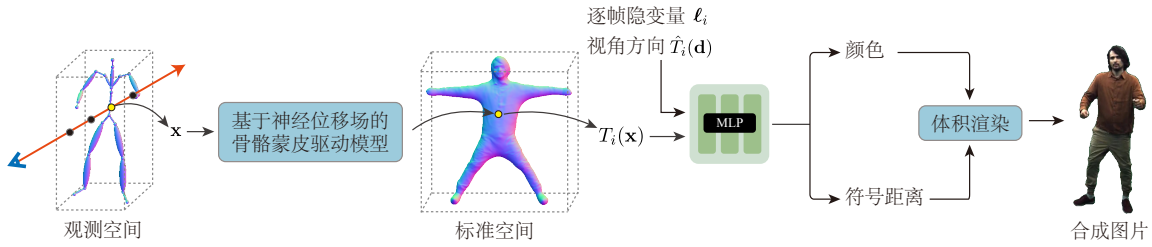


图 5-2 为了合成目标视角下的图片，本工作在观测空间发射相机射线并沿着射线采点，然后把采样点变换到标准空间，将其输入到 MLP 网络中以预测符号距离和颜色，最后通过体积渲染得到图片。

网格模型。其次，本章设计了一个基于隐式神经表示的位移场，并将其与基于参数化人体模型的骨骼蒙皮驱动算法结合，用于生成可建模衣物运动的空间变形场。相比于 Animatable NeRF 的变形场，本章提出的变形场更易于使用。最后，本章通过充分的消融实验证明了符号距离场和位移场的有效性，并在多个数据集上验证了本章提出的方法。实验结果表明，本方法在几何建模和新姿态合成两方面远远超过了之前的方法。

5.2 方法

5.2.1 方法概述

本文旨在从少量相机拍摄的视频中重建具有高质量几何的动态人体，并能在给定新的人体动作的情况下生成自由视点视频。本章实验假定相机已经同步，并且已知相机参数。类似于之前的研究工作[86,250]，本章假设已知图片中人体的掩模和三维人体姿态。本工作使用人体掩模将背景图像的像素值设置为零。图 5-2 展示了本章提出的模型。本模型将人体的几何形状和外观表示为标准空间中的符号距离场和颜色场 (第 5.2.2 节)，并使用位移场建模输入视频中的动态人体 (第 5.2.3 节)。第 5.2.4 节描述了如何使用体积渲染从 RGB 视频中学习动态的符号距离场。

5.2.2 动态人体模型

受到之前研究工作[82,86,248] 的启发，为了从输入视频中重构人体模型，本方法使用一个标准空间下的人体模型和一个变形场来表示视频中的动态人体。给定第 i 个视频帧的观测空间中的三维点 $\mathbf{x}$，变形场 T_i 将三维点 $\mathbf{x}$ 变换为标准空间，然后将其输入到标准空间下人体模型网络以预测其符号距离和颜色。

与之前使用密度场的方法[82,86,248]不同，本方法使用符号距离场[14,150]表示人体几何形状。给定标准空间中的任意三维点，本方法使用一个 MLP 网络预测其符号距离 s。本章将标准空间中几何模型记为 F_s。第 i 个视频帧中的动态人体几何模型被定义为：

$$(s_i(\mathbf{x}), \mathbf{z}_i(\mathbf{x})) = F_s(T_i(\mathbf{x})), \tag{5-1}$$

其中 $\mathbf{z}_i(\mathbf{x})$ 是网络输出的特征向量，作为后续颜色网络的输入。F_s 是一个具有九层全连接层的 MLP 网络。

与之前方法 IDR[150]类似，本章提出的颜色网络将空间位置、法向量和视角方向为输入。本方法还定义了逐帧的隐变量 $\boldsymbol{\ell}_i$，作为颜色网络的输入，以编码第 i 帧的场景状态。将标准空间中的颜色模型记为 $F_{\mathbf{c}}$，第 i 帧观测空间的颜色模型被定义为：

$$\mathbf{c}_i(\mathbf{x}) = F_{\mathbf{c}}(T_i(\mathbf{x}), \mathbf{z}_i(\mathbf{x}), \mathbf{n}_i(\mathbf{x}), \mathbf{d}, \boldsymbol{\ell}_i), \tag{5-2}$$

其中法向量 $\mathbf{n}_i(\mathbf{x})$ 是在三维点 $T_i(\mathbf{x})$ 处符号距离 $s_i(\mathbf{x})$ 的梯度。后面将描述用于变换空间点的变形场 T_i。$F_{\mathbf{c}}$ 是一个具有五层全连接层的 MLP 网络。对于新的人体姿势，本方法使用第一帧的隐变量 $\boldsymbol{\ell}_0$ 作为颜色网络的输入。本方法的人体几何网络 F_g 和颜色网络 $F_{\mathbf{c}}$ 参照了 IDR[150]的网络结构。逐帧的场景编码 $\boldsymbol{\ell}_i$ 的维度为 128。

5.2.3 神经位移场

本方法将视频中的人体运动分解为铰链式的和非刚性的变形，并分别用骨骼蒙皮驱动模型[21,86]和基于隐式神经表示的位移场来表示，以建立观测空间与标准空间之间的对应关系。相比于第 4 章使用神经蒙皮权重场建模变形场，本章提出的神经位移场有两个优点。首先，位移向量的参数量少于蒙皮权重向量的参数量，因此相对来说易于训练。本章的实验结果表明位移场提升了渲染效果。其次，对于新的人体姿态，本章可以使用 MLP 网络根据输入的人体姿态直接预测位移场，而第 4 章中的模型需要额外优化新人体姿态下的神经蒙皮权重场。

具体而言，对于第 i 个视频帧中的观测空间的三维点 $\mathbf{x}$，本方法根据输入的三维人体姿态计算 K 个变换矩阵 $G_i^k \in SE(3)$。基于骨骼蒙皮驱动算法[21,86]，本方法可以使用以下公式将该三维点变换到标准空间：

$$\bar{\mathbf{x}}' = \left(\sum_{k=1}^{K} w_i^k(\mathbf{x}) G_i^k \right)^{-1} \bar{\mathbf{x}}, \tag{5-3}$$

其中，$\bar{\mathbf{x}}$ 和 $\bar{\mathbf{x}}'$ 是 $\mathbf{x}$ 和 $\mathbf{x}'$ 的齐次坐标，$w_i^k(\mathbf{x})$ 是第 k 个人体部位的蒙皮权重。给定变换后的点 $\mathbf{x}'$，本方法使用神经位移场将其变形到人体表面上。记神经位移场为 $F_{\Delta\mathbf{x}}:(\mathbf{x},S_i)\rightarrow\Delta\mathbf{x}_i$，其中 S_i 是第 i 个视频帧的三维人体姿态参数。变形场 $T_i(\mathbf{x})$ 被定义为：

$$T_i(\mathbf{x})=\mathbf{x}'+F_{\Delta\mathbf{x}}(\mathbf{x}',S_i), \tag{5-4}$$

其中，$F_{\Delta\mathbf{x}}$ 是一个具有九层全连接层的 MLP 网络。在实现中，本方法基于 SMPL 网格模型[104] 得到骨骼蒙皮驱动算法所需的蒙皮权重。类似于之前的一些研究工作[60,86,252]，本方法首先找到输入的三维点 $\mathbf{x}$ 在 SMPL 网格模型上最近的网格顶点，然后将其蒙皮权重作为三维点 $\mathbf{x}$ 的蒙皮权重 $\mathbf{w}_i(\mathbf{x})$。请注意，本方法也可以使用其他参数化人体模型[108-109,251]来得到蒙皮权重。

5.2.4 模型训练

为了从输入的 RGB 视频中优化本章所提出的模型，本工作使用可微分渲染器将其渲染成图片，并最小化渲染图片和观察图片之间的差异。受到之前方法[17-18] 的启发，本方法适用 VolSDF[17] 中的体积渲染方案来渲染本章提出的动态符号距离场。给定第 i 帧的图像像素，本工作首先沿着其相机设想 $\mathbf{r}$ 在近边界和远边界之间采样 N_k 个点 $\{\mathbf{x}_k\}_{k=1}^{N_k}$。然后，本方法在这些三维点处预测符号距离和颜色。为了进行体积渲染，本方法根据下面的公式将符号距离 $s_i(\mathbf{x}_k)$ 转换为体素密度：

$$\sigma_i(\mathbf{x})=\begin{cases}\frac{1}{\beta}\left(1-\frac{1}{2}\exp\left(\frac{s_i(\mathbf{x})}{\beta}\right)\right) & \text{if } s_i(\mathbf{x})<0,\\ \frac{1}{2\beta}\exp\left(-\frac{s_i(\mathbf{x})}{\beta}\right) & \text{if } s_i(\mathbf{x})\geq 0,\end{cases} \tag{5-5}$$

其中 β 是一个可学习参数。最终，本方法使用数值积分计算出目标像素的渲染颜色 $\tilde{\mathbf{C}}_i(\mathbf{r})$：

$$\tilde{\mathbf{C}}_i(\mathbf{r})=\sum_{k=1}^{N_k}\alpha_i(\mathbf{x}_k)\prod_{j<k}(1-\alpha_i(\mathbf{x}_j))\mathbf{c}_i(\mathbf{x}_k), \tag{5-6}$$

其中，$\alpha_i(\mathbf{x}_k)=1-\exp(-\sigma_i(\mathbf{x}_k)\delta_k)$，$\delta_k$ 是相邻采样点 $\mathbf{x}_{k+1}$ 和 $\mathbf{x}_k$ 之间的距离。在本章的所有实验中，采样点的数量 N_k 都被设置为 64。

目标函数：本方法通过最小化渲染像素颜色与观测像素颜色之间的差异来优化模型参数。此外，本方法还使用掩模损失函数和程函方程（Eikonal equation）[258]作为监督信

号。本方法还为神经位移场添加了一个正则项。具体而言，渲染误差函数被定义为：

$$L_{\text{rgb}} = \sum_{r \in \mathcal{R}} \|\tilde{\mathbf{C}}_i(\mathbf{r}) - \mathbf{C}_i(\mathbf{r})\|_2, \tag{5-7}$$

其中 $\mathbf{C}_i(\mathbf{r})$ 是第 i 帧的真实像素颜色，$\mathcal{R}$ 是相机射线集合。为了用掩模监督符号距离场，本方法找到相机射线 $\mathbf{r}$ 上最小的符号距离 $s_i^{\mathbf{r}}$，并施加交叉熵损失函数 BCE：

$$L_{\text{mask}} = \sum_{r \in \mathcal{R}} \text{BCE}(\text{sigmoid}(-\rho s_i^{\mathbf{r}}), M_i(\mathbf{r})), \tag{5-8}$$

其中 $M_i(\mathbf{r}) \in 0, 1$ 是真实掩模值。类似于 IDR[150]，本工作将 ρ 设置为 50，并在每 10000 次迭代时将其乘以 2。在实验中，ρ 更新 5 次后保持为常数。

本方法在观察空间中采样一组三维点 $\mathcal{X}_i$，并对这些采样点施加程函损失函数：

$$L_{\text{E}} = \sum_{\mathbf{x} \in \mathcal{X}_i} (\|\nabla F_s(T_i(\mathbf{x}))\|_2 - 1)^2. \tag{5-9}$$

为了正则化位移场，本方法还在标准空间中采样了一组三维点 $\mathcal{X}_i'$，然后计算正则项：

$$L_{\Delta\mathbf{x}} = \sum_{\mathbf{x} \in \mathcal{X}_i'} \|F_{\Delta\mathbf{x}}(\mathbf{x}, S_i)\|_2. \tag{5-10}$$

用于训练的总目标函数为：

$$L = L_{\text{rgb}} + L_{\text{mask}} + \lambda_1 L_{\text{E}} + \lambda_2 L_{\Delta\mathbf{x}}, \tag{5-11}$$

其中，λ_1 被设置为 0.1，λ_2 被设置为 0.01。本方法采用 Adam 优化器[240]训练模型参数，其学习率从 $5e^{-4}$ 开始，并随着训练的进行指数衰减，最终衰减为 $5e^{-5}$。

5.3 实验分析

本文评估了两种可驱动人体表征方式：（1）基于神经位移场和体素密度场的数字人体模型 NeRF-PDF；（2）基于神经位移场和符号距离场的数字人体模型 SDF-PDF。实验中，NeRF-NBW 代表第 4 章中提出的基于神经蒙皮权重场和体素密度场的数字人体模型。本章采用 IDR 网络结构构建了 NeRF-NBW 模型，因此实验结果与第 4 章中的结果有所不同。本章没有评估基于神经蒙皮权重场和符号距离场的数字人模型，因为本工作在实验中发现该表征方式在优化过程中容易陷入局部最小值。

5.3.1 数据集和实验指标

Human3.6M 数据集[23]：该数据集包含了 4 个摄像机拍摄的多视角视频，使用基于标记的运动捕捉系统收集了人类姿态。和第 4 章的实验相同，本工作使用三个相机进行训练，剩余的相机用于测试。

MonoCap 数据集[35,219]：为了充分进行实验，本章收集了一个多视角数据集，其中的多视角视频包括了两个来自 DeepCap 数据集[35] 和两个来自 DynaCap 数据集[219] 的视频。这两个数据集都使用稠密视角的相机拍摄动态人体，并提供了人类掩码和三维人类姿态。本章选择一个相机视角训练模型，并选择十个均匀分布的相机视角进行测试。本工作选择每个视频的一个片段来进行实验。每个片段有 300 帧用于训练，并有 300 帧用于评估新人体姿态合成。

ZJU-MoCap 数据集[250]：第 3 章中使用多视角相机阵列创建了该数据集，其中包含 9 个多视角视频。本章按照第 3 章中的实验设定选择四个均匀分布的相机视角作为训练输入，并在剩余的视角上进行测试。为了进一步探索本章提出的方法对于整合时序信息的能力，本工作还在第一个相机视角上训练模型，并在剩余视角上进行测试。

SyntheticHuman 数据集：为了测试重建结果，本章创建了一个合成数据集，其中包含 7 个动态的三维人体模型。4 个人体模型在保持 A-姿势的同时进行旋转，其被渲染成单目视频。另外 3 个人体模型执行随机的动作，其被渲染成四个视角的视频。所有视频帧和相机视角都用于模型的训练。该数据集仅用于评估三维人体几何重建的性能。

实验指标：对于三维几何重建，本工作参考 PIFu[56] 使用两个指标：点到表面的欧氏距离（Point-to-surface Euclidean distance，P2S）和倒角距离（Chamfer distance，CD）。这两个指标的单位为厘米。对于图像合成，本工作遵循 NeRF[15] 使用两个指标来评估各个方法：峰值信噪比（PSNR）和结构相似性指数（SSIM）。

5.3.2 图片合成的实验结果

Human3.6M 数据集上的结果：表 5-1 和 5-2 将本方法与其他方法[25,248,250,259]在图像合成上进行了量化比较。在新视角合成和新人体姿态合成方面，本章提出的 NeRF-PDF 和 SDF-PDF 都优于基线方法。图 5-3 和图 5-4 展示了本方法和基线方法在训练姿态和新姿势下的新视角合成的定性比较。本方法产生了逼真的渲染结果，大大超过了基线方

表 5-1 Human3.6M 数据集上的训练人体姿态的新视角合成结果。“NeRF-PDF” 和 “SDF-PDF” 是本章提出的两个数字人模型，“NeRF-NBW” 是第 4 章中提出的人体模型。这些方法在三个视角视频上训练，然后在剩余的一个视角上测试。

	S1	S5	S6	S7	S8	S9	S11	平均
指标	PSNR ↑							
D-NeRF[248]	19.63	20.92	20.64	17.90	20.81	23.79	17.23	20.13
NB[250]	22.87	24.60	22.82	23.17	21.72	24.28	23.70	23.31
NHR[25]	21.08	20.64	20.40	20.29	19.13	23.04	21.91	20.93
A-NeRF[259]	23.09	24.32	23.81	24.10	22.25	25.50	24.60	23.95
NeRF-NBW	24.37	24.01	24.19	23.98	23.66	25.71	25.31	24.46
NeRF-PDF	**24.43**	**24.80**	24.37	**24.57**	23.61	**26.03**	25.43	**24.75**
SDF-PDF	24.41	24.42	**24.39**	24.34	**23.87**	25.87	**25.65**	24.71
指标	SSIM ↑							
D-NeRF[248]	0.838	0.807	0.811	0.722	0.845	0.889	0.737	0.807
NB[250]	0.897	**0.917**	0.888	0.914	0.894	0.910	0.896	0.903
NHR[25]	0.872	0.872	0.830	0.868	0.871	0.879	0.871	0.866
A-NeRF[259]	0.905	0.914	0.888	**0.915**	0.902	0.913	0.902	0.906
NeRF-NBW	0.900	0.897	0.885	0.903	0.908	0.908	0.908	0.901
NeRF-PDF	0.903	0.915	0.889	0.911	0.909	0.917	0.904	0.907
SDF-PDF	**0.915**	0.915	**0.899**	**0.915**	**0.915**	**0.920**	**0.920**	**0.914**

法。虽然第 3 章中提出的模型在训练过的人体姿态上合成了高质量的图像，但在新人体姿态上则难以给出合理的渲染结果。相比之下，本方法通过使用基于骨骼蒙皮驱动模型的变形场来驱动人体模型，在图像生成过程中具有更好的可控性，这与经典的图形学管线类似。本方法在训练过的人体姿态的新视角合成上也得到了最好的效果。

MonoCap 数据集上的结果：表 5-3 和 5-4 总结了本方法和其他方法[25,250,259] 在训练人体姿态和新人体姿势下进行新视角合成的定量比较。本方法在视角合成方面的表现均优于基线方法。图 5-4 展示了本方法和基线方法在 MonoCap 数据集上进行图像合成的定性比较。实验结果显示本模型能够在只有一个输入视角的情况下合成高质量的图片。

表 5-2 Human3.6M 数据集上的新人体姿态合成结果。

	S1	S5	S6	S7	S8	S9	S11	平均
指标	PSNR ↑							
NB[250]	22.11	**23.51**	23.52	22.33	20.94	23.04	23.72	22.74
NHR[25]	20.48	20.72	20.47	19.66	18.83	22.18	22.12	20.64
A-NeRF[259]	21.14	22.72	22.89	22.58	21.35	23.97	23.14	22.54
NeRF-NBW	23.00	23.17	**24.72**	**23.14**	**22.64**	24.36	24.52	23.65
NeRF-PDF	**23.51**	23.45	24.47	23.09	22.50	**24.78**	**24.63**	**23.78**
SDF-PDF	23.08	23.50	24.31	23.09	22.64	24.45	**24.61**	23.67
指标	SSIM ↑							
NB[250]	0.879	0.897	0.889	0.889	0.876	0.884	0.884	0.885
NHR[25]	0.853	0.860	0.856	0.852	0.855	0.860	0.867	0.858
A-NeRF[259]	0.872	0.885	0.882	0.888	0.883	0.888	0.875	0.882
NeRF-NBW	0.885	0.881	0.900	0.887	0.895	0.885	0.895	0.890
NeRF-PDF	0.891	0.892	0.890	0.886	0.894	0.894	0.896	0.892
SDF-PDF	**0.893**	**0.901**	**0.903**	**0.893**	**0.899**	**0.898**	**0.907**	**0.899**

ZJU-MoCap 数据集上的结果：表 5-5 列出了量化结果。当使用 4 个相机视角训练模型时，本方法在训练过的人体姿势的新视角合成方面的性能与第 3 章中提出的模型相当，并且在新姿态合成方面优于该模型。图 5-3 和 5-4 展示了在 4 个相机视图上训练的模型的定性结果。当仅使用单个相机视图进行训练时，本方法在训练人体姿态和新人体姿态上都显著优于第 3 章中提出的模型。

5.3.3 三维几何重建的实验结果

为了验证本章提出的方法在三维几何重建方面的性能，本章在 SyntheticHuman 数据集、Human3.6M 数据集、MonoCap 数据集上与之前的方法[248,250,259]进行了比较。由于 NHR[25]没有重建人体几何结构，因此本工作没有与 NHR 进行比较。本工作使用 Marching Cubes 算法[241]从神经隐式表示中提取人体的三维几何模型。对于使用体素密度场的方

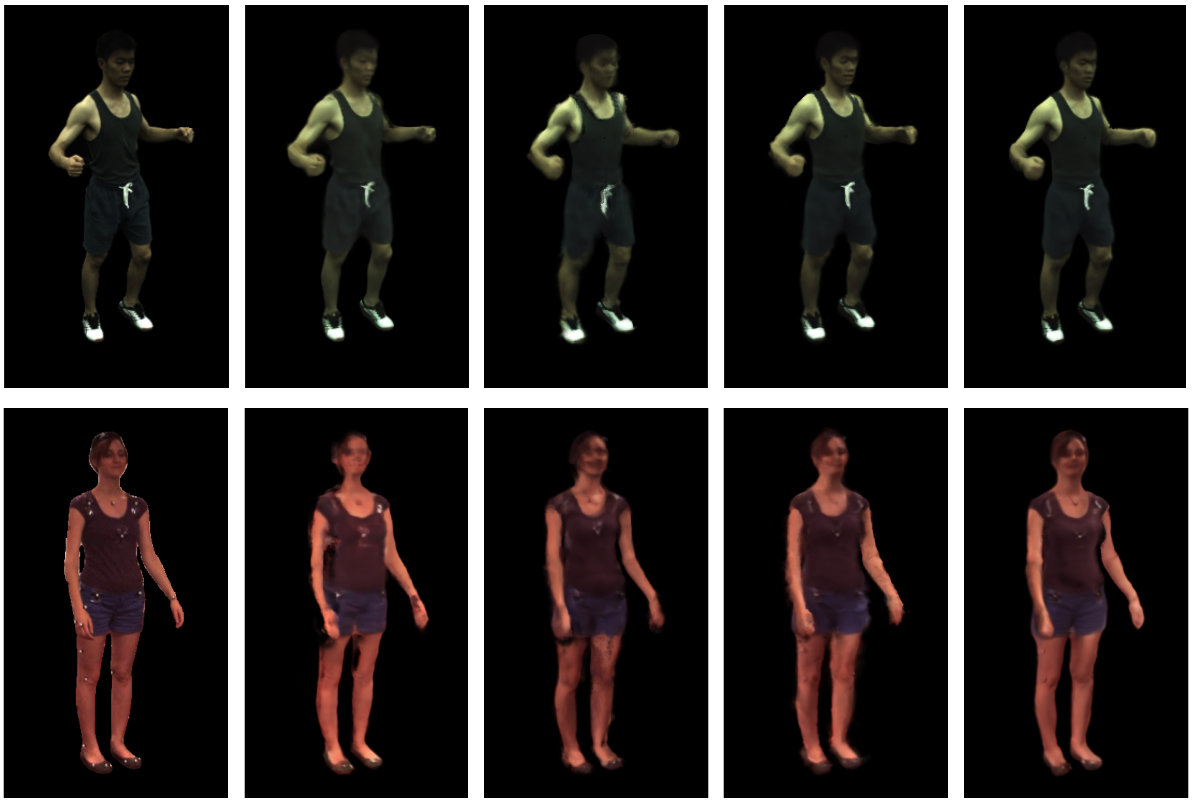


图 5-3 ZJU-MoCap 和 Human3.6M 数据集上的训练姿态下的新视角合成结果。第一行和第二行分别是 ZJU-MoCap 数据集和 Human3.6M 数据集的结果。“NeRF-PDF” 和 “SDF-PDF” 是本章提出的人体表征方法。“Neural Body” 和 “NeRF-NBW” 分别是第 3 章和第 4 章中提出的数字人模型。

表 5-3 MonoCap 数据集上的训练人体姿态的新视角合成结果。“NeRF-PDF” 和 “SDF-PDF” 是本章提出的两个数字人模型，“NeRF-NBW” 是第 4 章中提出的人体模型。这些方法在一个视角视频上训练，然后在其他的十个或十一个视角上测试。

	Lan	Marc	Olek	Vlad	平均	Lan	Marc	Olek	Vlad	平均
指标	PSNR ↑					SSIM ↑				
NB[250]	22.63	24.60	22.41	17.40	21.76	0.883	0.883	0.867	0.854	0.872
NHR[25]	21.36	22.93	23.21	**17.65**	21.29	0.857	0.878	**0.893**	**0.871**	0.875
A-NeRF[259]	21.93	22.99	21.64	15.51	20.52	0.871	0.868	0.846	0.798	0.845
NeRF-NBW	22.14	23.75	22.60	17.37	21.47	0.873	0.882	0.870	0.850	0.868
NeRF-PDF	**23.47**	**24.84**	**23.54**	17.52	**22.34**	**0.890**	0.904	0.883	0.854	0.883
SDF-PDF	22.41	24.37	23.21	17.57	21.89	0.887	**0.905**	0.885	0.862	**0.885**

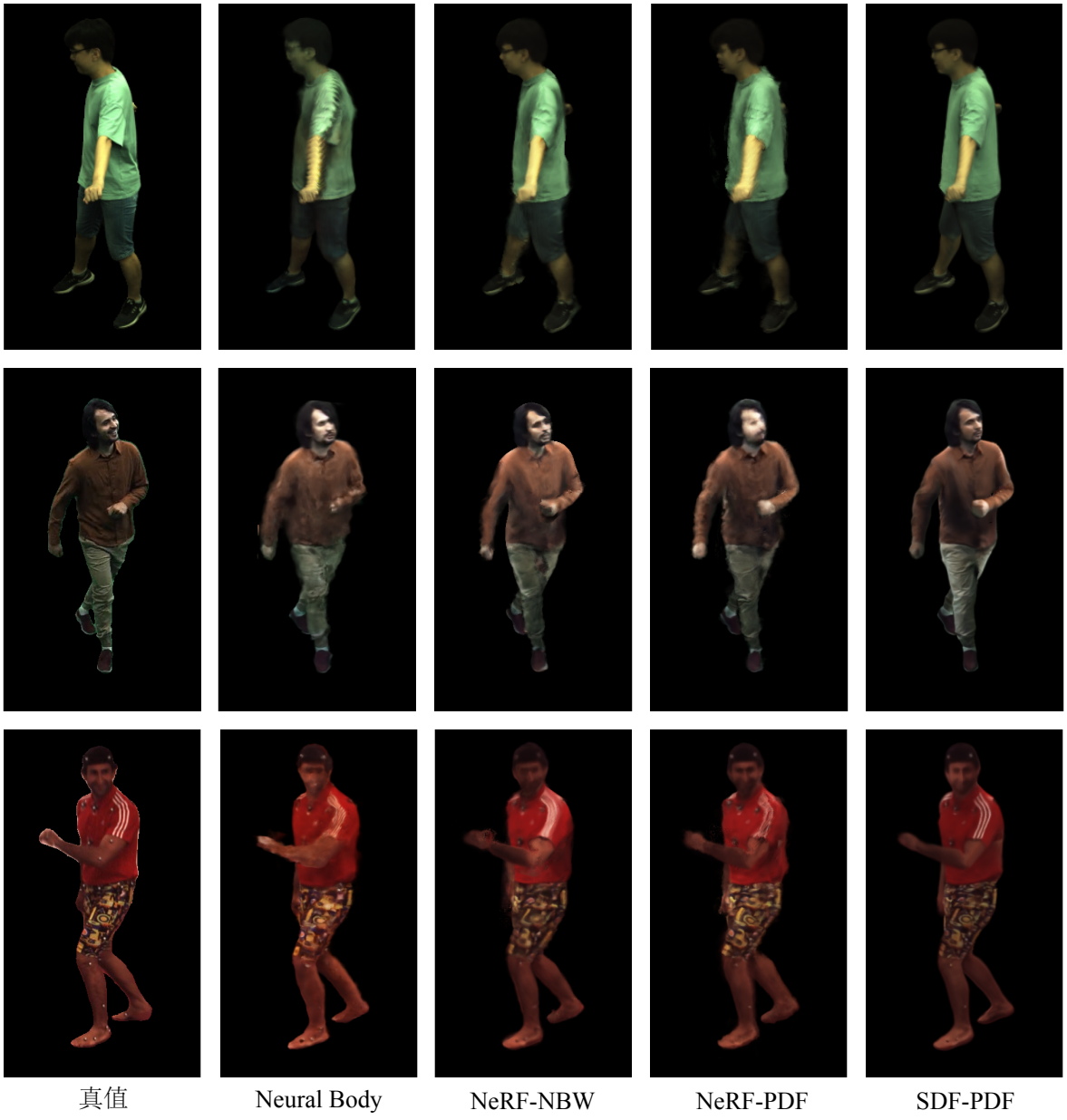


图 5-4 ZJU-MoCap、MonoCap 和 Human3.6M 数据集上的新人体姿态合成结果。第一行是 ZJU-MoCap 数据集上的结果。第二行是 MonoCap 数据集上的结果。最后一行是 Human3.6M 数据集上的结果。“Neural Body” 和 “NeRF-NBW” 分别是第 3 章和第 4 章中提出的数字人模型。

法，本工作在应用 Marching Cubes 算法时经验性地设置阈值来提取几何形状。对于本章中提出的人体表示模型 “SDF-PDF”，本工作将阈值设置为零。

SyntheticHuman 数据集的结果：表 5-6 和 5-7 比较了本方法与其他方法[86,250,259]在 P2S 和 CD 指标方面的表现。之前的方法[86,250,259]、第 4 章提出的方法 “NeRF-NBW”、本章提出的方法 “NeRF-PDF” 都使用体素密度场来表示人体三维几何，而本章提出的模型 “SDF-PDF” 使用符号距离场建模人体几何。实验结果表明模型 “SDF-PDF” 比其他

表 5-4 MonoCap 数据集上的新人体姿态合成结果。

	Lan	Marc	Olek	Vlad	平均	Lan	Marc	Olek	Vlad	平均
指标	PSNR ↑					SSIM ↑				
NB[250]	21.77	22.91	21.13	17.49	20.83	0.867	0.868	0.836	0.848	0.854
NHR[25]	20.49	21.64	**21.96**	**17.69**	20.45	0.843	0.875	**0.875**	**0.872**	0.866
A-NeRF[259]	21.27	21.41	20.09	15.35	19.53	0.855	0.853	0.808	0.795	0.828
NeRF-NBW	21.93	22.03	21.36	17.31	20.66	0.869	0.875	0.848	0.847	0.860
NeRF-PDF	**22.52**	**23.18**	21.59	17.46	**21.19**	**0.874**	0.889	0.852	0.850	0.866
SDF-PDF	21.72	22.62	21.63	17.56	20.88	0.872	**0.890**	0.856	0.858	**0.869**

表 5-5 ZJU-MoCap 数据集上的新视角合成结果。“NeRF-PDF” 和 “SDF-PDF” 是本章提出的两个数字人模型，“NeRF-NBW” 是第 4 章中提出的人体模型。

	4 个视角				单个视角			
	训练人体姿态		新人体姿态		训练人体姿态		新人体姿态	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NB[250]	**28.15**	**0.943**	24.05	0.896	20.90	0.802	20.32	0.797
NeRF-NBW	25.69	0.913	23.60	0.893	23.00	0.879	22.62	0.871
NeRF-PDF	27.50	0.935	**24.35**	0.904	**23.86**	0.894	**22.98**	0.883
SDF-PDF	27.45	0.940	24.28	**0.909**	23.80	**0.901**	22.92	**0.889**

方法至少高出 0.72 P2S 和 0.63 CD。图 5-5 展示了定性比较的结果。

真实数据的结果：为了进一步验证本方法的有效性，本工作还在真实数据集上进行了重建实验。由于真实数据上没有人体三维几何模型的真值，因此本章只呈现了定性比较的结果。图 5-6 展示了在 Human3.6M 和 MonoCap 数据集上的重建结果。这些方法在 MonoCap 数据集的一个视角和 Human3.6M 数据集的三个视角上进行了训练。实验结果表明，模型 “SDF-PDF” 能够从稀疏视角视频中重建出高质量的人体三维几何。

表 5-6 SyntheticHuman 数据集上的三维几何重建结果。前四行显示的是在单目视频上重建的结果，其余行显示的是在 4 个视角视频上重建的结果。A-NeRF[259] 在视频 “S5” 上无法收敛。

	P2S ↓					
	D-NeRF[248]	NB[250]	A-NeRF[259]	NeRF-NBW	NeRF-PDF	SDF-PDF
S1	3.49	1.44	1.30	4.30	1.59	**0.75**
S2	3.38	1.68	1.39	4.66	1.74	**0.70**
S3	3.96	1.52	1.80	4.45	1.61	**0.62**
S4	4.18	1.20	1.46	2.84	1.58	**0.58**
S5	1.22	1.20	37.5	2.87	1.85	**0.66**
S6	1.76	1.31	1.17	2.62	1.97	**0.74**
S7	1.66	1.61	1.03	2.88	2.11	**0.69**
平均	2.81	1.42	6.52	3.52	1.78	**0.70**

表 5-7 SyntheticHuman 数据集上的三维几何重建结果。前四行显示的是在单目视频上重建的结果，其余行显示的是在 4 个视角视频上重建的结果。A-NeRF[259] 在视频 “S5” 上无法收敛。

	CD ↓					
	D-NeRF[248]	NB[250]	A-NeRF[259]	NeRF-NBW	NeRF-PDF	SDF-PDF
S1	2.40	1.39	1.29	2.94	1.45	**0.86**
S2	2.45	1.48	1.22	3.03	1.51	**0.81**
S3	2.71	1.42	1.53	3.02	1.40	**0.81**
S4	2.85	1.23	1.28	2.07	1.40	**0.74**
S5	1.10	1.14	36.4	2.33	1.44	**0.65**
S6	1.43	1.28	1.07	2.05	1.54	**0.73**
S7	1.82	1.74	1.29	2.59	2.02	**0.65**
平均	2.11	1.38	6.30	2.57	1.54	**0.75**

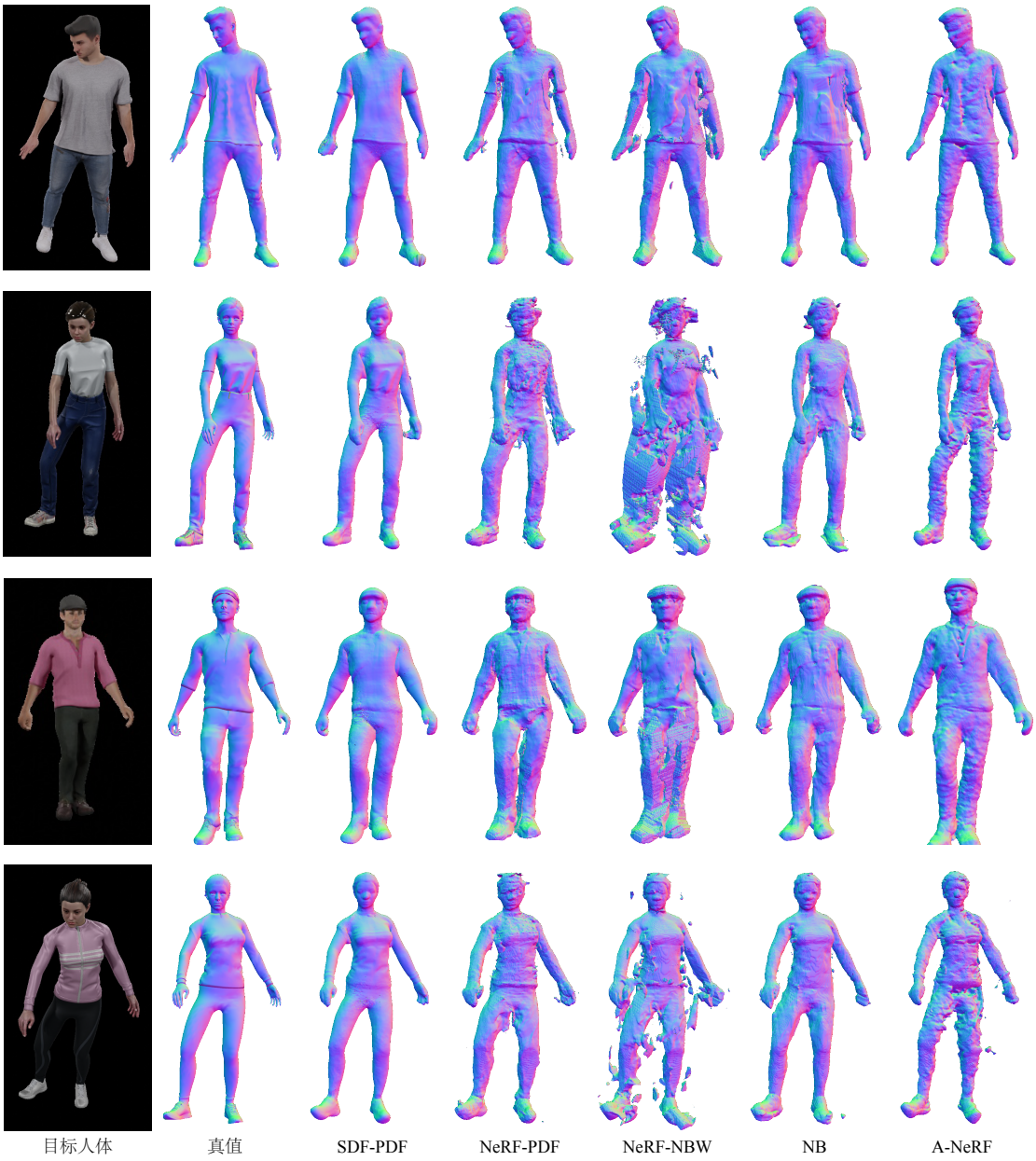


图 5-5 SyntheticHuman 数据集上的三维人体几何重建结果。第一行的结果是在 4 个视角视频上的重建结果，第二行的结果是在单目视频上的重建结果。本方法 “SDF-PDF” 显著优于其他方法。

5.3.4 消融实验

本节进行了消融实验来分析本工作的设计选择和训练数据如何影响模型的性能。

符号距离场的影响：表 5-1、5-2、5-3、5-4、5-5 显示，“NeRF-PDF” 和 “SDF-PDF” 在 Human3.6M、MonoCap 和 ZJU-MoCap 数据集上的图像合成方面性能相近。但是，在三维重建方面，符号距离场相比体素密度场取得了更好的重建结果。在表 5-6 和 5-7 中，“SDF-PDF” 在 SyntheticHuman 数据集的三维重建方面显著优于 “NeRF-PDF”。图 5-5 和

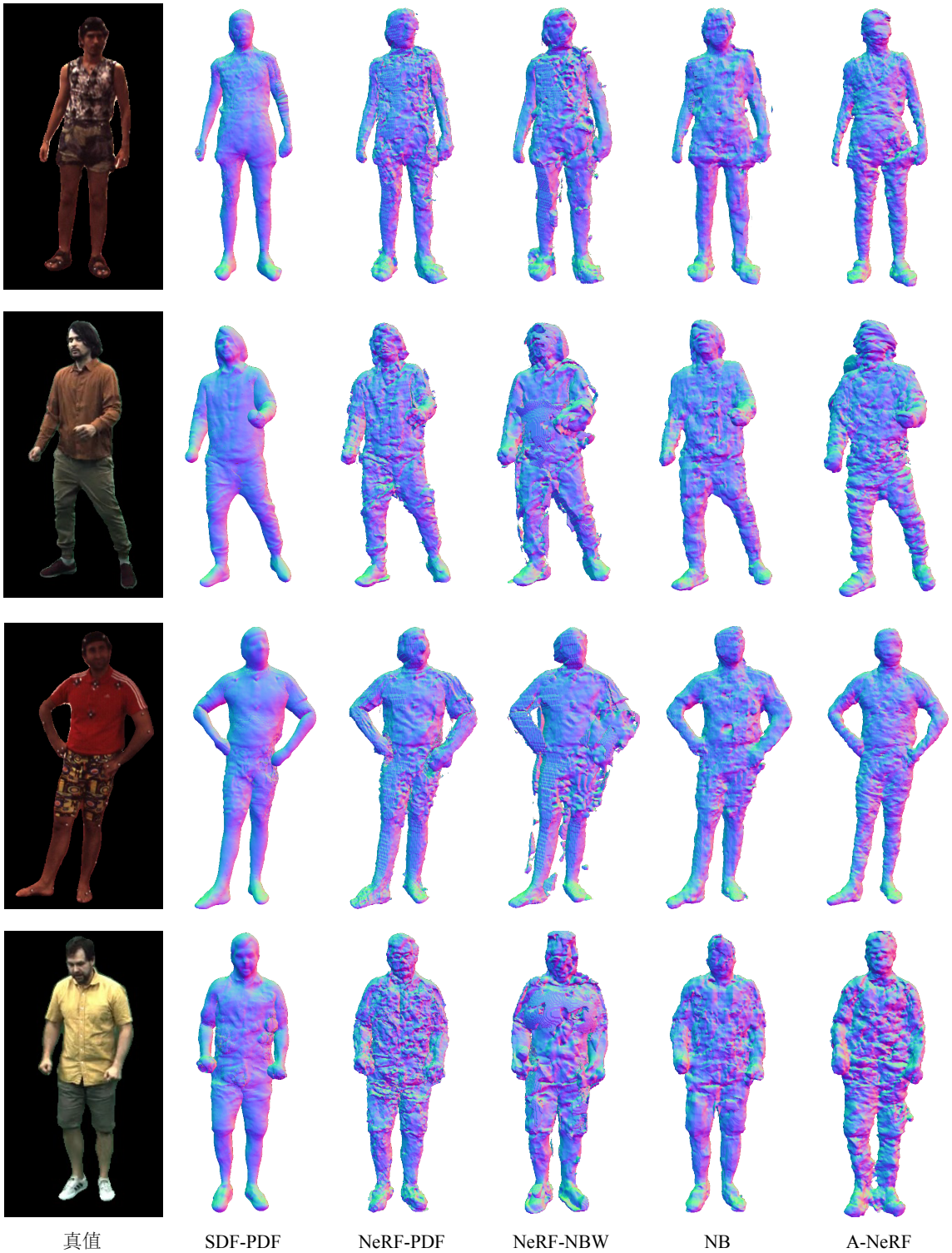


图 5-6 Human3.6M 和 MonoCap 数据集上的三维人体几何重建结果。第一行的结果是在三个视角的视频上的重建结果，第二行的结果是在单视角视频上的重建结果。本方法 “SDF-PDF” 明显优于其他方法。如图所示，第 4 章中提出的模型 “NeRF-NBW” 的重建结果较为粗糙。

5-6 的定性结果也表明，“SDF-PDF” 的重建结果更好。

神经位移场的影响：在 Human3.6M 数据集上的新姿态合成结果显示，“NeRF-NBW”

和 “NeRF-PDF” 的表现分别为 23.65 和 23.78，具有相近的性能。然而，在 MonoCap、ZJU-MoCap 和 SyntheticHuman 数据集上，“NeRF-NBW” 在图像合成和三维几何重建方面的表现都比 “NeRF-PDF” 差。原因可能是这三个数据集的人体姿态是基于无标记姿态估计系统从图像中估计的，因此可能不太准确。因为神经蒙皮权重场完全将人体运动建模为基于骨骼驱动的变形，所以它对姿态精度的敏感性比基于神经位移场的方法更高。

网络结构的影响：本章中的模型 “NeRF-NBW” 采用了 IDR[150]的网络，而第 4 章中的模型使用了 NeRF[15]的网络。新采用的网络具有比 NeRF 网络更大的颜色 MLP 网络。在 Human3.6M 数据集上，第 4 章中的原始网络在新姿态合成方面得到了 22.55 PSNR，而新采用的网络提供了 23.65 PSNR，表明更大的网络提高了渲染性能。

5.4 总结与讨论

本章提出了一个基于符号距离场的动态人体几何表示方法，用于从稀疏视角的视频中重建出高质量的三维人体几何。与第 4 章中的方法类似，本章提出的模型将动态人体表示为变形场和标准空间下的人体模型。与第 4 章中的方法不同的是，本章将人体模型表示为符号距离场和颜色场。除此之外，本章还提出了基于神经位移场的骨架蒙皮驱动框架，更好地建模了人体的动态形变。本章使用了基于符号距离场的体积渲染技术，从稀疏视角的视频中优化得到了动态人体表示。在 Human3.6M、MonoCap、ZJU-MoCap 和 SyntheticHuman 数据集上的实验结果表明，本章所提出的模型在三维重建和图像合成方面相比之前的方法取得了更好的效果。

第 6 章 基于多层感知机图的动态场景表示

6.1 引言

本章的第 3、4、5 章节分别提出了一个神经人体表示，致力于从视频中重建出高质量的动态三维人体模型，用于合成动态人体的体积视频，从而支持用户以任意视角观看目标动态场景，给用户带来沉浸式的观看体验。然而，前几章提出的神经人体模型推理成本较高，无法实现实时渲染。支持实时渲染的体积视频有很多重要的应用，比如沉浸式视频会议、体育赛事的自由视角观看、远程教育。除此之外，体积视频还需要能够被有效地压缩，以减小视频的存储空间并支持高效通信传输。设计一个满足这些要求的体积视频表示是一个仍未解决的问题。

传统的体积视频方法大多采用基于图像直接进行新视角合成。这些研究工作利用光场插值技术（Light field interpolation）[11,260] 或者图像变形技术（Image warping）[261-262] 构建自由视点视频系统。此类工作使用许多同步的摄像机记录动态场景，然后基于输入的视图的插值来合成新的视图。这些方法本质上是在使用稠密的多视角视频构建体积视频。虽然已经有很多的多视角视频编码技术，但稠密多视角视频的存储空间仍然很大，从而使得传输多视角视频需要较多时间，难以满足实时观看视频的需求。另一类研究工作[7,9] 利用 RGB-D 传感器重建带纹理的三维网格模型以作为动态场景的表示。通过网格压缩技术，这种动态场景表示可以非常紧凑，并能够支持流媒体形式的体积视频。但这些方法仅能在受限的环境中捕获人类和物体，因为重建通用动态场景的可渲染网格模型仍然是一个非常具有挑战性的问题。

近期神经场景表示的发展[26,81,88] 为体积视频的表示提供了一种新的解决方案。这些研究工作使用神经网络表示三维场景，可以通过可微分的渲染器从图像中有效地学习神经网络参数。例如，Neural Volumes[81] 使用由三维卷积神经网络预测的一组三维体素网格来表示体积视频，其中每个网格存放相应的颜色和体素密度。该工作基于体积渲染[15] 将三维体素网格投影到图片空间，并通过最小化渲染误差来优化三维卷积神经网络。虽然该工作取得了较好的效果，但由于三维体素网格容易消耗大量 GPU 内存，因此难以通过网络预测高分辨率的三维体素网格，也因此难以建模高分辨率的三维场景。神经辐射场 NeRF[15] 通过用全连接神经网络回归任意三维点的体素密度和颜色来表示连续的

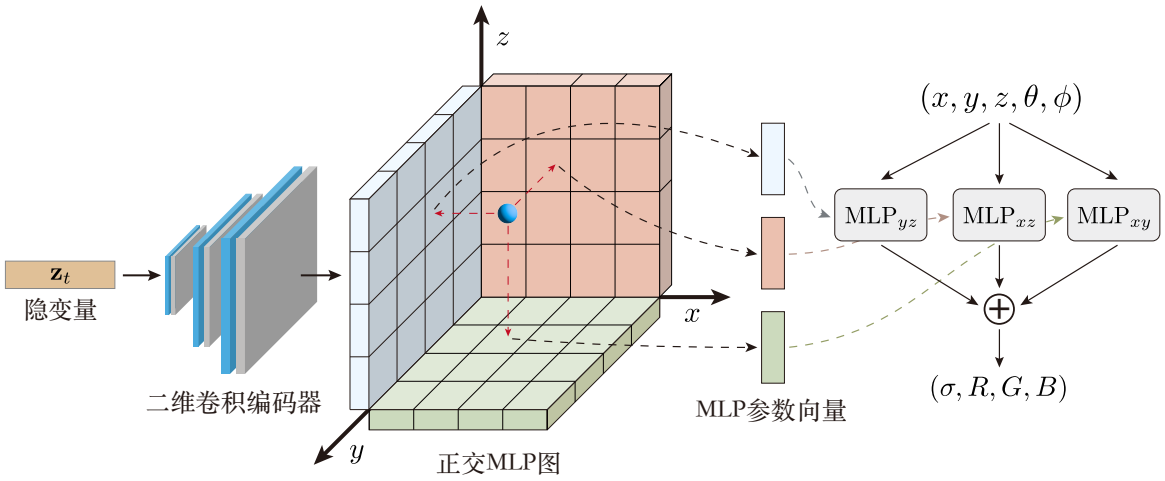


图 6-1 动态 MLP 图的基本思想。与之前的动态渲染方法 DyNeRF[26]使用大型 MLP 网络不同，本章提出的模型利用二维卷积神经网络来动态地生成每一帧的二维 MLP 图，其中每个像素存储着一个小型 MLP 网络的参数向量。通过这种方式，本章可以将体积视频表示为一组小型 MLP 网络，从而减少了网络推理成本，大大提高了渲染速度。

三维场景，从而理论上能够合成任意分辨率的图像。受到该特性的启发，DyNeRF[26] 基于神经辐射场开发了一种体积视频表示。该工作通过在多层感知机（MLP）的输入中加入时变隐变量，使得神经辐射场可以表示不同时刻的三维场景。尽管神经辐射场的渲染质量很高，但由于神经网络的推理成本高，且渲染所需推理次数较多，该工作的渲染速度通常非常慢。为了提高渲染速度，一些方法[71,73,88] 预先计算神经辐射场的值，并将其缓存在一个高效的数据结构中，如八叉树（Octree）。这种策略通常导致较高的存储成本。对于静态场景，只需存储一帧，所以存储成本是可以接受的。但对于动态场景，每一帧存储一个缓存模型将导致极高的存储成本。

本章提出一种新的体积视频表示，称为动态 MLP 图，用于动态场景的高质量实时渲染。图 6-1 展示了本工作的基本思想。相对于神经辐射场 NeRF 用单个 MLP 网络对体积视频进行建模，本方法使用一组小型 MLP 网络建模三维场景，并且使用一个二维卷积解码器基于逐帧的隐变量预测这些 MLP 网络的参数。具体来说，输入一个多视图视频，本方法选择其中的一部分视角，并将这些视角图片输入二维卷积编码器中以获得每帧的隐变量。然后，本方法适用一个二维卷积解码器从隐变量中回归得到一个二维网格，该网格中的每个像素存储着一个 MLP 网络的参数向量。本章称这样的二维网格为 MLP 图。为了使用 MLP 图对三维场景进行建模，本方法首先将三维空间中的任意点投影到 MLP 图上，然后获取相应的 MLP 网络参数，最后将网络参数加载进 MLP 网络以

推断三维点的体素密度和颜色。用许多小型的 MLP 网络表示三维场景可以降低网络推理的成本，从而增加三维场景的渲染速度。之前的工作[74,168] 已经提出这样的策略，但是这些工作为每个静态场景显式地存储了网络参数。对于动态场景，显式存储网络参数很容易占据大量的存储空间。相比之下，本工作使用二维卷积解码器作为一个超网络，用于高效地预测每一个视频帧的 MLP 网络参数，从而有效地在时间域上压缩存储空间。本章所提出的动态 MLP 图的另一个优势是 MLP 图使用二维网格表示三维场景，从而可以通过二维卷积神经网络作为解码器。相比于 Neural Volumes[81] 工作中使用的三维卷积神经网络，二维卷积神经网络的推理速度更快，而且占用了更少的显存。

综上所述，本章的主要贡献可以概括为以下几点：首先，本章提出了一种名为动态 MLP 图的体积视频表示。本方法通过二维卷积神经网络预测动态 MLP 图，从而可以建模存储空间较小的体积视频，并且支持快速的网络推理。其次，本章基于动态 MLP 图构建了一个实时渲染动态场景的新管线。最后，本章在 NHR[25] 和 ZJU-MoCap[83] 数据集上验证了动态 MLP 图的有效性。在所有数据集中，本章提出的方法在渲染质量和速度方面表现出最先进的性能，同时占用了较少的存储空间。本章还通过充分的消融实验证明了本方法每个策略设计的有效性。

6.2 方法

给定一个由同步相机拍摄的多视角视频，本章工作的任务是生成一个高质量、占用磁盘存储空间较小、支持实时渲染的体积视频。本章提出了一种称为动态 MLP 图的新颖的隐式神经表示方法，并且开发了一个动态场景的实时视角合成流程。本节首先描述如何使用 MLP 图建模三维场景（第 6.2.3 节）。接下来，第 6.2.2 节讨论如何使用动态 MLP 图表示体积视频。最后，第 6.2.3 节介绍了一些加速渲染过程的策略。

6.2.1 基于多层感知机图的三维场景建模

MLP 图是一个二维网格图，该网格的每个像素存储了一个 MLP 网络的参数向量。为了使用 MLP 图表示三维静态场景，本方法将空间中的任意三维点 $\mathbf{p}$ 投影到一个二维平面上以查询相应的 MLP 参数。该平面由 MLP 图定义。在实际实现中，本方法将 MLP 图与坐标系的轴对齐，因此三维点 $\mathbf{p}$ 会被正交投影到一个坐标平面上。图 6-2（b）展示

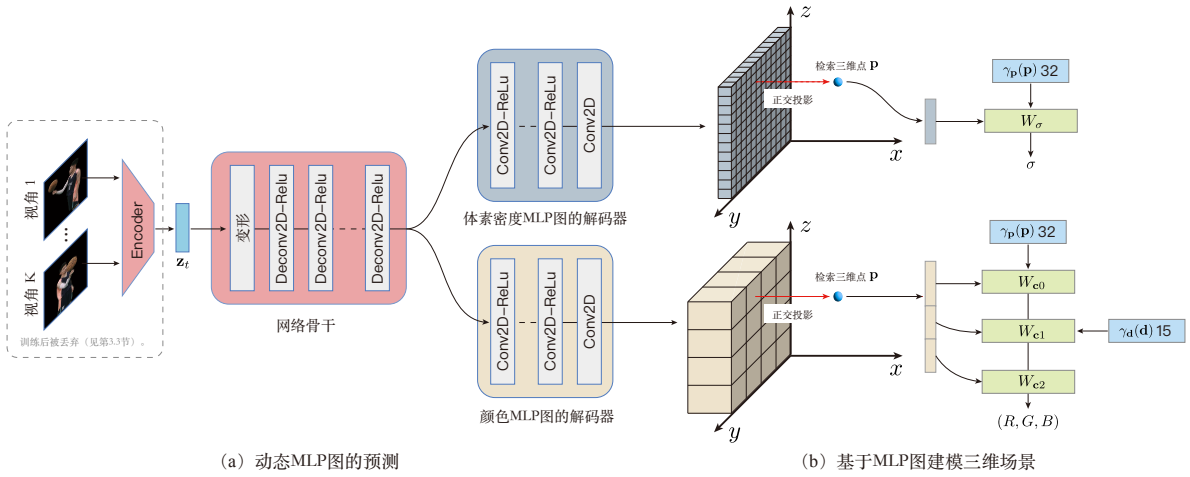


图 6-2 定义在 YZ 平面上的动态 MLP 图的示意图。本方法使用所有视频帧共享的自编码器网络预测 MLP 图以表示每一帧视频中的三维场景。（a）具体来说，对于一个特定的视频帧，编码器网络将输入视角的一部分图片映射到一个隐向量中，然后该向量被解码器网络处理以输出体素密度 MLP 图和颜色 MLP 图。（b）对于任意三维查询点，本方法将其投影到 MLP 图所定义的二维平面上，然后从该平面上检索对应的网络参数以构建体素密度 MLP 网络和颜色 MLP 网络。最后，构建得到的 MLP 网络预测该点的体素密度和颜色。

了一个例子，其中 MLP 图定义在 YZ 平面上。投影后得到的二维点通过空间分块被分配到 MLP 图上的某个像素，随后该像素中的参数向量被动态加载到 MLP 网络中。本方法采用了一个小型的 MLP 网络来预测查询点 $\mathbf{p}$ 的密度和颜色。该 MLP 网络包含了一个用于预测密度的一层全连接网络和一个用于预测颜色的三层全连接网络。MLP 图中所有像素上的 MLP 网络共同建模目标三维场景。由于每个 MLP 网络只描述目标场景的一部分区域，所以本章提出的模型可以使用较小的 MLP 网络实现高质量渲染。之前的一些研究工作[74,168,263-264]也使用了一组网络表示三维场景。相比于这些工作，本章所提出的 MLP 图采用 2D 平面的形式，因此本方法能够使用二维卷积神经网络有效且高效地生成 MLP 参数，这将在下文中详细描述。

在预测三维空间点 $\mathbf{p}$ 的体素密度和颜色时，为了提升网络的建模能力，本方法并没有直接将空间坐标点输入到 MLP 网络中，而是将输入的坐标点映射到高维的特征向量。具体而言，本方法定义了三个多层的哈希表[77]：$\mathbf{h}_{xy}$，$\mathbf{h}_{xz}$，$\mathbf{h}_{yz}$。每个哈希表的分辨率为 $L \times T \times F$，其中 L 是哈希表的层数，T 是表格大小，F 是特征维度。为了映射输入的三维坐标点 $\mathbf{p}$ 到高维特征向量，本方法首先将其投影到与三个轴对齐的正交平面上，然后使用三个多层哈希表将投影点转换为特征向量，最后通过求和聚合三个特征向量。映

射过程的公式被定义为：

$$\gamma_{\mathbf{p}}^{h}(\mathbf{p}) = \eta(x, y, t; \mathbf{h}_{xy}) + \eta(x, z, t; \mathbf{h}_{xz}) + \eta(y, z, t; \mathbf{h}_{yz}), \tag{6-1}$$

其中 (x, y, z) 为三维点 $\mathbf{p}$ 的坐标，t 为视频的帧数，η 为编码函数。该编码函数根据输入点从哈希表中获取特征向量（详见 Instant-NGP[77]）。此外，本工作借鉴 EG3D[202]使用二维卷积神经网络预测三面体特征图（Tri-plane feature map），并使用三面体特征图为每个三维空间点分配一个高维特征向量 $\gamma_{\mathbf{p}}^{t}(\mathbf{p})$。本方法将哈希表特征向量 $\gamma_{\mathbf{p}}^{h}(\mathbf{p})$ 和三面体特征向量 $\gamma_{\mathbf{p}}^{t}(\mathbf{p})$ 相加，以获得最终的特征向量 $\gamma_{\mathbf{p}}(\mathbf{p})$。体素密度 σ 和颜色 $\mathbf{c}$ 通过以下方式进行预测：

$$(\sigma, \mathbf{c}) = M(\gamma_{\mathbf{p}}(\mathbf{p}), \gamma_{\mathbf{d}}(\mathbf{d})), \tag{6-2}$$

其中，M 表示 MLP 网络，$\gamma_{\mathbf{d}}(\mathbf{d})$ 是经过编码的视角方向。图 6-2（b）可视化了 MLP 图建模三维场景的过程。在实际实现中，本方法将 $\gamma_{\mathbf{d}}$ 实现为一个位置编码函数[15]，而 2D CNN 则与用于动态 MLP 映射的 2D CNN 共享参数，其详细结构将在第 6.3 节中介绍。

相互正交的 MLP 图：本章的实验表明，只使用定义在一个平面上的 MLP 图来建模三维场景难以获得较好的渲染质量。原因是目标场景的内容可能是沿着某个坐标轴的高频信号函数，这使得 MLP 图难以拟合场景内容。在这种情况下，场景内容可能沿着其他坐标轴有着较低的频率。受到之前研究工作[78,202,238]的启发，本方法在三个与坐标轴对齐的正交平面上定义了三个 MLP 面。当预测查询点的值时，本方法首先使用 MLP 图来预测体素密度和颜色 $\{(\sigma_i, \mathbf{c}_i) \,|i = 1, 2, 3\}$，然后通过求和对这些值进行聚合。图 6-1 可视化了求和正交信号函数的输出的过程。第 6.4.2 节的实验结果表明，定义在正交平面上的三个 MLP 图的渲染性能优于定义在相同平面上的三个 MLP 图。

6.2.2 基于动态多层感知机图的体积视频表示

基于 MLP 图，本方法可以使用二维卷积神经网络来表示体积视频。给定一个多视角视频，本方法适用二维卷积神经网络回归每个视频帧的包含一组 MLP 参数的 2D 网格图，这些网络参数编码了该时刻下的三维场景的几何形状和外观。图 6-2（a）展示了这个过程。本方法采用了一个编码器-解码器作为网络架构，其中编码器从输入的相机视角回归隐变量，而解码器根据隐变量生成 MLP 图。

具体而言，对于特定的视频帧，本方法选择一些相机视角作为网络的输入，并使用二维卷积编码器将这些视角转换为一个隐变量。这个编码过程和 Neural Volumes[81] 中的过程相同。编码器输出的隐变量用于记录相应的视频帧的场景状态，并用于预测 MLP 图。另一种获取隐变量的方式是预定义每个视频帧的可学习特征向量[14,265-266]。使用编码器网络的优点在于编码器网络隐式地在不同视频帧之间共享信息，从而能在训练过程中利用时序观测信息联合重建目标场景，而不仅仅是逐帧的重建。

给定编码器输出的隐变量，本方法使用二维卷积解码器来预测 MLP 图。图 6-2（a）给出了二维卷积解码器的结构。本章将隐变量记为 $\mathbf{z} \in \mathbb{R}^D$。本方法首先使用一个全连接网络将 $\mathbf{z}$ 映射到一个 4096 维的特征向量，然后将此特征向量重排为一个带有 256 个通道的 4×4 特征图。接下来，一个带有一系列反卷积层的网络将其上采样为更高分辨率的 $D \times D$ 特征图。基于这个特征图，本方法使用两个的卷积网络分别预测体素密度和颜色的 MLP 图。这两个卷积网络由多个卷积层组成。通过控制卷积层的数量和步长，本方法可以控制网络输出的 MLP 图的分辨率。由于体素密度 MLP 网络的参数比颜色 MLP 网络少，因此本方法可以为体素密度 MLP 图预测更高的分辨率。这样能取得更好的渲染性能。第 6.4.2 节的实验结果验证了这一策略的有效性。

6.2.3 加速渲染过程

本章提出的模型使用一组小型的 MLP 网络来表示三维场景。由于这些 MLP 网络比 DyNeRF[26]的网络要小得多，因此本方法的网络评估所需的时间更少，从而使得本方法有更快的渲染速度。为了进一步提高渲染速度，本章引入了两个额外的策略。

首先，在训练完成之后，本工作丢弃了编码器网络。具体而言，本工作使用已训练的编码器为每个视频帧计算相应的隐变量并存储。这样避免了每次渲染过程都要重新进行编码器网络的前向推理，从而节省了编码器的推理时间。

其次，本工作通过跳过三维空间中的空区域来减少网络推理的次数。为了实现这一策略，本工作为每个视频帧计算一个低分辨率的三维占据体（3D occupancy volume），其中每个体素存储了一个二进制值，用于指示该体素是否被占据。该占据体从已训练的网络模型中提取而来。当一个体素的密度高于一个阈值时，本工作将该体素中占据值设为 1。因为占据体的分辨率较低且以二进制格式存储，所以一个 300 帧的视频的占据体只占用大约 8MB 的存储空间。在渲染过程中，本工作只在占据值为 1 的空间区域中进行

网络推理。在实际实现中，本工作先计算三维点的体素密度，然后再计算三维点的颜色，以此来进一步减少颜色网络的推理次数。具体而言，本工作首先计算在占据值为 1 的体素中的采样点的体素密度，然后基于这些体素密度计算体积渲染公式中的权重值[15]。如果某个三维点的权重高于一定阈值，那么本工作预测其颜色值，否则直接跳过该点。

6.3 实现细节

网络结构：本章提出的模型采用了一个具有七个卷积层的编码器。类似于 Neural Volumes[81]的方法，这个编码器将三个分辨率为 512×512 的图像作为输入，并将这些图片转换成一个 256 维的隐变量。解码器网络由一个主干网络和两个预测器组成。主干网络有六个反卷积层，用于将输入的 4×4 特征图上采样到 256×256 的主干特征图。本方法使用一个卷积层从主干特征图回归到一个 96 通道的特征图，并将其变形为三个具有 32 通道的特征图。体素密度 MLP 图的预测器是基于步幅为 1 的卷积层实现的。该预测器输出一个分辨率为 256×256 的 MLP 图，图中每个像素的 MLP 网络有 32 个参数（32×1）。颜色 MLP 图的预测器将四个步幅为 2 的卷积层和一个步幅为 1 的卷积层应用于输入特征图，得到一个分辨率为 16×16 的 MLP 图。图中每个像素存放的 MLP 网络有 2624 个参数（$32 \times 32 + (32 + 15) \times 32 + 32 \times 3$）。MLP 图中的 MLP 网络使用 ReLU 作为激活函数。对于每个多层哈希表的超参设置，本方法遵循 Instant-NGP[77]的做法，设置 $L = 19$，$T = 16$ 和 $F = 2$。

存储空间：对于一个 300 帧的视频，本模型中隐变量、MLP 图解码器、多层哈希表和占据体的存储成本分别为 300 KB、103MB、131MB 和 8MB。请注意，因为本方法在第 6.2.3 节中采用的加速策略丢弃了编码器网络，所以不需要存储编码器网络。

训练细节：用于训练模型的目标函数被定义为：

$$L = L_c + \lambda_{KL} L_{KL}. \tag{6-3}$$

L_c 是观测像素颜色 $\tilde{C}(\mathbf{r})$ 和渲染像素颜色 $C(\mathbf{r})$ 的误差：

$$L_c = \sum_{\mathbf{r} \in \mathcal{R}} ||\tilde{C}(\mathbf{r}) - C(\mathbf{r})||_2^2, \tag{6-4}$$

其中 $\mathcal{R}$ 表示相机射线的集合。Kullback-Leibler 散度损失 L_{KL} 的计算方式遵循 Neural Volumes[81]。在本章所有的实验中，本文将 λ_{KL} 设为 $1e^{-6}$。如果目标场景仅包含前景物

表 6-1 各个模块对渲染质量的贡献。量化结果是模型在 ZJU-MoCap 和 NHR 数据集上的两个场景的渲染结果平均值。

方法	基线方法 1	C-NeRF	本方法-单 MLP	本方法
描述	三平面特征图 + 单个 MLP 网络	特征体 + 单个 MLP 网络	三平面特征图 + 哈希表 + 单个 MLP 网络	三平面特征图 + 哈希表 + 动态 MLP 图
LPIPS	0.072	0.076	0.065	**0.058**

体，本方法还会增加掩模损失函数（Mask loss）来帮助训练。该函数测量了渲染图像不透明度 $M(\mathbf{r})$ 与真实前景掩模 $\tilde{M}(\mathbf{r})$ 之间的误差：

$$L_m = \sum_{\mathbf{r}\in\mathcal{R}} ||\tilde{M}(\mathbf{r}) - M(\mathbf{r})||_2^2. \tag{6-5}$$

实验中将掩模损失函数的权重设为 0.1。在优化过程中，本方法将批量大小（Batch size）设置为 8，在一张 RTX 3090 GPU 上训练模型。训练过程耗时约 16 小时。网络和多层哈希表的学习率分别被设置为 $5e^{-4}$ 和 $5e^{-3}$，并在训练中指数衰减为原先值的 0.1 倍。

6.4 实验分析

6.4.1 数据集

为了评估本章提出的模型的性能，本工作在 ZJU-MoCap[83]和 NHR[25]数据集上进行实验。这两个数据集使用多个同步相机捕获前景动态人体，其拍摄的动态场景展现了较大的运动幅度。在 ZJU-MoCap 数据集上，本章实验均匀选择了 11 个摄像机进行训练，并使用其余视图进行评估。所有视频序列的长度为 300 帧。在 NHR 数据集上，90%的相机被用作训练视图，其他视图用于评估。本文选择从 NHR 数据集的视频中提取 100 帧用于重建每个场景的体积视频。这两个数据集为前景动态场景提供了前景的掩模。

6.4.2 消融实验

本章在 ZJU-MoCap 和 NHR 数据集中分别选择了一个场景进行消融实验，用于验证本模型的设计选择。表 6-1 和表 6-2 给出了消融实验的定量结果。

表 6-2 ZJU-MoCap 和 NHR 数据集上的消融实验。本实验测试了各个模型在一块 RTX 3090 GPU 上渲染一张 512×512 分辨率的图像的时间。第 6.4.2 节详细描述了表中每一行的模型。

	LPIPS ↓	渲染时间 (毫秒) ↓	存储 (MB) ↓
（1）颜色 MLP 图分辨率 1×1	0.069	22	489
（2）颜色 MLP 图分辨率 4×4	0.065	23	324
（3）颜色 MLP 图分辨率 32×32	0.060	27	208
（4）体素密度 MLP 图分辨率 16×16	0.064	25	250
（5）单个 MLP	0.065	90	133
（6）单个 MLP w/o ESS	0.065	1345	**125**
（7）单个 XY 平面上的 MLP 图	0.072	**19**	206
（8）w/o 正交 MLP 图	0.067	24	242
（9）w/o ESS	**0.058**	144	237
（10）完整模型	**0.058**	24	242

哈希表和 MLP 图对渲染性能的影响：表 6-1 列出了消融实验的结果。表中模型 “基线方法 1” 由三平面特征图（Tri-plane feature map）和一个被所有视频帧共享的 MLP 网络组成。该 MLP 网络具有与 NeRF 相同的网络。模型 “本方法-单 MLP” 是表 6-2 中行（5）的模型。表 6-1 的实验结果表明了渲染质量中的两个重要模块：（1）哈希表（Ours-Single vs. C-NeRF[267]）。（2）动态 MLP 图（Ours vs. Ours-Single）。使用动态 MLP 图使得每个 MLP 网络只需要表示一个小的场景区域。相比于用单个 NeRF 编码整个体积视频，MLP 图用一组 MLP 网络更好地拟合了目标动态场景。

MLP 分辨率的影响：表 6-2 的第 1、2、4 行的实验结果表明，降低分辨率会削弱渲染性能。请注意，分辨率降低时存储成本会增加。这是因为本方法在下采样时添加了更多的卷积层，如第 6.3 节所述。表 6-2 第 3 行的结果表明，增加分辨率并不一定会提高渲染质量。一个合理的解释是，增加分辨率会使模型更难以训练。这有两个原因。首先，更高的 MLP 图分辨率意味着需要优化更多的参数。其次，由于显存的限制，增加分辨率需要在每个训练迭代中减少输入到每个颜色 MLP 中的查询点数，从而降低了批量大

图 6-3 ZJU-MoCap 数据集上不同 MLP 图分辨率的模型的定性结果。四个模型的颜色 MLP 图分辨率为 1×1、4×4、16×16 和 32×32。

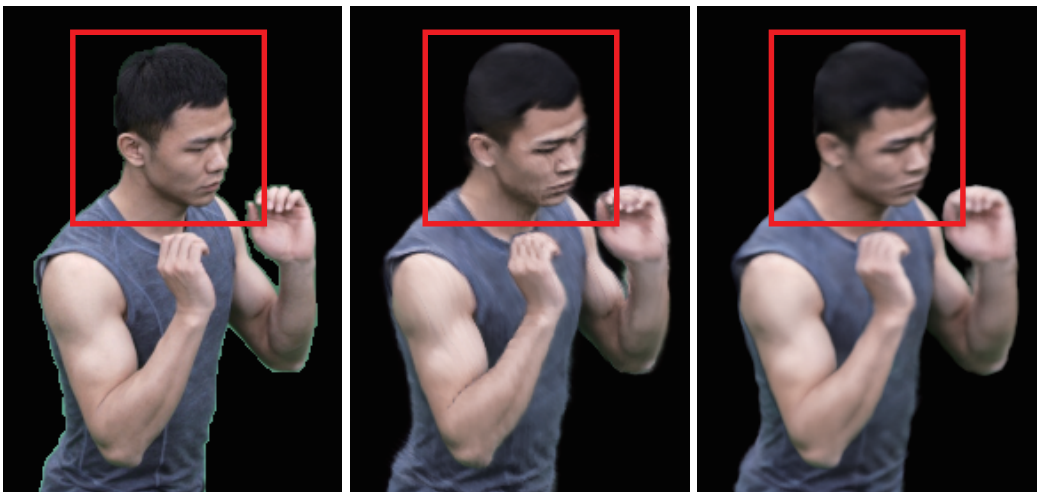


图 6-4 NHR 数据集上动态 MLP 图和单个 MLP 网络的比较。

小。图 6-3 呈现了一些定性结果，可以看到，分辨率为 16×16 的 MLP 图取得了最好的渲染质量。表 6-2 的第 1-4 行结果还表明了 MLP 图的分辨率并不会对渲染速度产生太大影响。这可以归因于 MLP 网络的推理次数与 MLP 图的分辨率无关。

之所以密度 MLP 图的分辨率可以达到 256×256 而颜色 MLP 图只达到了 16×16，原因在于密度 MLP 网络只有 32 个参数，而颜色 MLP 网络具有 2624 个参数。这意味着 256×256 的密度 MLP 图的参数数量与 28×28 颜色 MLP 图相近。此外，场景几何通常比外观具有更低的频率。因此，场景几何相比外观更容易被网络拟合。

动态 MLP 图的影响：表 6-2 的第 5-6 行中的消融实验将动态 MLP 图替换为被所有视频帧共享的单个 MLP 网络。该 MLP 网络的实现方式与 NeRF 网络[15]相同，区别只在于该网络的输入是第 6.2.2 节中的特征向量 $\gamma_{\mathbf{p}}(\mathbf{p})$。实验结果表明，相比基于单个共

图 6-5 NHR 数据集上正交 MLP 图的消融实验。实验结果表明，正交 MLP 图可以提高渲染质量。

享 MLP 网络的模型而言，基于动态 MLP 图的模型具有更好的渲染质量和更快的渲染速度。第 6 行的结果表明，在跳过空区域的推理后，“单个 MLP” 模型的渲染速度显著增加，但仍然比本章提出的完整模型慢得多。图 6-4 呈现了可视化对比。

正交 MLP 图的影响：表 6-2 中的 “单个 XY 平面上的 MLP 图” 模型使用单个动态 MLP 图编码三维场景，该动态 MLP 图定义在 XY 平面上。实验结果表明，该模型渲染质量大幅度下降。模型 “w/o 正交 MLP 图” 采用了三个动态 MLP 图编码场景，这些 MLP 图都定义在 XY 平面上。该模型也未能取得跟本章提出的完整模型同样的渲染质量。

各模块对渲染时间的影响：表 6-2 中的 “w/o ESS” 模型没有跳过空区域的网络计算。实验结果表明，跳过空区域的网络计算可以将渲染速度加速 6 倍，而仅增加 5 MB 的存储。本消融实验还分析了本方法中各模块的运行时间。MLP 图解码器和查询哈希表的推理时间分别为 4ms 和 7ms。颜色 MLP 图和密度 MLP 图的总推理时间为 12ms。

6.4.3 和基线方法的比较

基线方法：本章将提出的方法与四个之前的动态渲染方法进行比较。（1）Neural Volumes（NV）[81]使用三维卷积网络生成 RGBα 体素网格作为体积视频。（2）C-NeRF[267]是 Neural Volumes 的后续工作，采用了三维卷积网络生成特征体，并使用一个 MLP 网络来预测辐射场。（3）D-NeRF[84]将动态场景分解为变形场和标准空间中的静态场景。（4）DyNeRF[26]使用一个 MLP 网络来表示体积视频，该网络以空间坐标和时间隐变量作为输入，并预测相应的视频帧的辐射场。由于 C-NeRF 和 DyNeRF 没有公开代码，本章重新实现了这两个工作以进行比较。

实验结果：表 6-3、6-4、6-5、6-6 列出了本方法与之前的方法[26,81,84,267]在渲染质量、

表 6-3 NHR 数据集上的量化结果。量化结果是 NHR 数据集上所有场景的平均值。

	Sport1	Sport2	Sport3	Basketball	平均
指标	PSNR ↑				
Neural Volumes[81]	31.76	31.48	31.04	29.17	30.86
C-NeRF[267]	31.81	32.12	31.99	**29.35**	31.32
D-NeRF[84]	30.12	30.18	29.66	27.02	29.25
DyNeRF[26]	31.76	32.43	31.33	27.97	30.87
本方法	**32.92**	**33.19**	**33.59**	29.11	**32.20**
指标	SSIM ↑				
Neural Volumes[81]	0.951	0.933	0.940	0.938	0.941
C-NeRF[267]	0.954	0.950	0.950	0.942	0.949
D-NeRF[84]	0.934	0.917	0.914	0.914	0.920
DyNeRF[26]	0.954	0.945	0.944	0.929	0.943
本方法	**0.959**	**0.954**	**0.956**	**0.943**	**0.953**
指标	LPIPS ↓				
Neural Volumes[81]	0.106	0.143	0.131	0.141	0.130
C-NeRF[267]	0.085	0.107	0.097	0.118	0.102
D-NeRF[84]	0.111	0.169	0.155	0.163	0.150
DyNeRF[26]	0.095	0.119	0.114	0.142	0.118
本方法	**0.067**	**0.084**	**0.076**	**0.094**	**0.080**

表 6-4 NHR 数据集上的平均渲染时间和存储。该数据集包含分辨率 512×612 和 384×512 的图像。

	NV[81]	C-NeRF[267]	D-NeRF[84]	DyNeRF[26]	本方法
渲染时间（毫秒）	73	1969	2303	5195	**33**
存储（MB）	658	1019	**4**	12	239

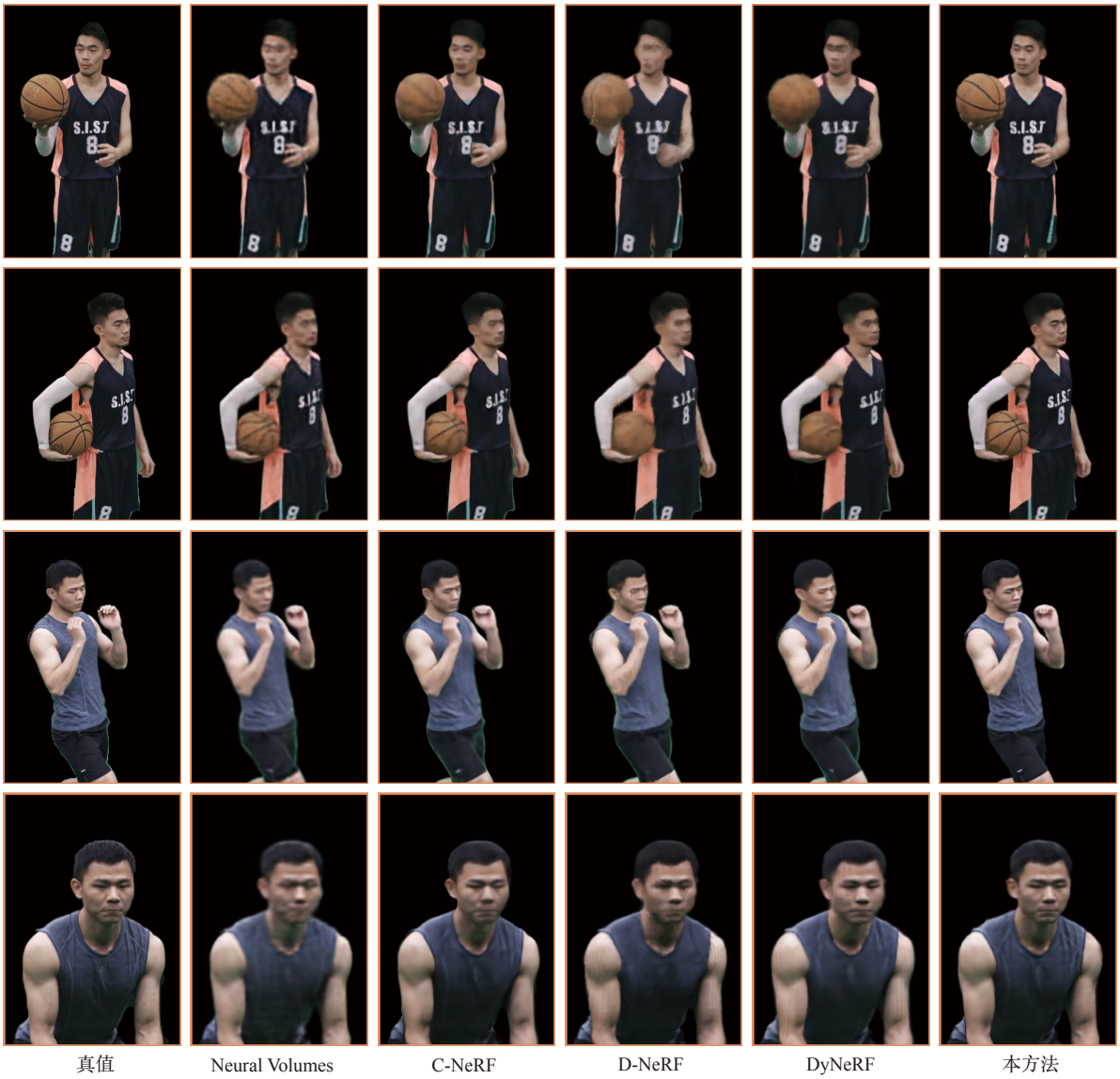


图 6-6 NHR 数据集上的定性比较。本章提出的方法在渲染质量上明显优于其他方法。第一行的结果表明，本方法可以合成篮球的纹理细节，而其他方法的渲染结果较为模糊。

渲染速度和存储方面的比较。本章采用 PSNR、SSIM 和 LPIPS[268]作为评估渲染质量的指标。在所有指标中，本章提出的模型在所有方法中取得了最好的表现。此外，本模型在保持较小的存储成本的同时，渲染速度比 C-NeRF、D-NeRF、DyNeRF 快两个数量级。因为 Neural Volumes 采用了显式的场景表示，所以渲染速度较快。本方法的渲染速度仅是 Neural Volumes 的渲染速度的两倍，但渲染精度显著提高。

图 6-6、6-7 展示了本方法和基准方法的定性结果。由于 Neural Volumes 的三维体素网格分辨率较低，其渲染结果较为模糊。当存在复杂的场景运动时，D-NeRF 的渲染结果也较为模糊，因为该方法难以在具有复杂运动的场景上学习变形场。由于 C-NeRF 和

表 6-5 ZJU-MoCap 数据集上的量化结果。量化结果是 ZJU-MoCap 数据集上所有场景的平均值。

	Twirl	Taichi	Warmup	Punch1	Punch2	Kick	Swing1	Swing2	Swing3	平均
指标	PSNR↑									
Neural Volumes[81]	30.23	26.92	26.90	30.15	25.84	28.06	28.76	28.28	27.95	28.12
C-NeRF[267]	31.84	29.03	28.92	30.75	27.52	29.81	30.82	29.81	29.62	29.79
D-NeRF[84]	27.48	26.68	26.20	28.78	25.53	28.10	27.37	27.47	26.14	27.08
DyNeRF[26]	31.50	**30.29**	28.92	30.88	**27.90**	**30.14**	30.09	29.88	29.28	29.88
本方法	**32.15**	29.94	**29.40**	**31.05**	27.89	30.10	**31.06**	**30.15**	**29.78**	**30.17**
指标	SSIM ↑									
Neural Volumes[81]	0.949	0.947	0.921	0.949	0.910	0.924	0.939	0.932	0.936	0.934
C-NeRF[267]	0.973	0.968	0.958	0.958	0.952	0.957	0.959	0.952	0.955	0.959
D-NeRF[84]	0.927	0.939	0.920	0.937	0.918	0.933	0.911	0.915	0.897	0.922
DyNeRF[26]	0.970	0.976	0.960	**0.960**	**0.953**	**0.959**	0.953	0.952	0.952	0.959
本方法	**0.976**	**0.977**	**0.963**	**0.960**	**0.953**	**0.959**	**0.962**	**0.957**	**0.958**	**0.963**
指标	LPIPS ↓									
Neural Volumes[81]	0.103	0.114	0.147	0.122	0.169	0.149	0.127	0.124	0.128	0.131
C-NeRF[267]	0.057	0.076	0.079	**0.072**	**0.080**	0.071	0.077	0.090	0.086	0.077
D-NeRF[84]	0.122	0.106	0.150	0.132	0.140	0.119	0.166	0.150	0.162	0.139
DyNeRF[26]	0.070	0.061	0.083	0.082	0.094	0.083	0.113	0.099	0.102	0.087
本方法	**0.049**	**0.047**	**0.069**	**0.072**	0.081	**0.068**	**0.074**	**0.072**	**0.080**	**0.068**

表 6-6 ZJU-MoCap 数据集上的平均渲染时间和存储。该数据集图像的分辨率为 512×512**。**

	NV[81]	C-NeRF[267]	D-NeRF[84]	DyNeRF[26]	本方法
渲染时间（毫秒）	49	1313	1534	3452	**24**
存储（MB）	658	1019	**4**	12	245

DyNeRF 使用了单个 MLP 网络拟合整个场景，这两个工作的渲染结果往往缺少高频细节。和这些方法相比，本方法取得了高质量的新视角合成结果。

图 6-7 ZJU-MoCap 数据集上的定性比较。相比于其他方法，本章提出的模型可以生成更加清晰的细节，如最后两行的结果所示。

6.5 总结与讨论

本章提出了一种新颖的隐式表示方法，称为动态 MLP 图，用于建模体积视频，实现动态场景的实时渲染。本方法的关键思想是利用共享的二维卷积神经网络预测每个视频帧的二维 MLP 图，这些 MLP 图的每个像素存储了一个 MLP 网络的参数。为了用 MLP 图建模三维场景，本方法将任意三维点投影到由 MLP 图定义的二维平面上，然后查询相应的 MLP 网络，最后回归该三维点的体素密度和颜色。本章提出使用二维卷积神经网络预测每一帧的二维 MLP 图，可以有效地减少存储成本。除此之外，本章预先计算了占据体，用于跳过空区域的计算，进一步加速了渲染速度。在 ZJU-MoCap 和 NHR 数据集上的实验结果表明，本章提出的方法在保持较小的低存储成本的情况下，实现了动态场景的实时渲染和取得了最好的渲染质量。

第 7 章 总结与展望

7.1 全文总结

本文围绕动态三维人体的建模与渲染展开研究，提出了一系列新颖的人体数字化技术，实现了从稀疏视角视频中重建高渲染质量、高几何精度、可驱动的数字人模型，并支持动态人体的实时渲染。为了从稀疏视角图片中构建高质量的人体模型，本文提出了基于结构化隐变量的动态人体表示，整合了输入视频的时序信息，从而获得足够的观测以优化得到正确的目标人体。针对数字人的可驱动性问题，本文提出神经蒙皮权重场，实现了骨骼蒙皮驱动算法与神经辐射场的结合。在此基础上，本文构建了基于符号距离场的动态人体几何表示，对几何的优化过程施加正则化，从而获得了高质量的三维人体几何。最后，本文提出使用 MLP 图表示动态场景，充分挖掘了小型 MLP 网络在渲染速度上的优势，并使用二维卷积神经网络高效地预测动态 MLP 图，搭建了面向动态场景的高真实感实时渲染管线。总体而言，本文的主要创新点有以下几个方面：

（1）为了实现从少数同步相机拍摄的多视角视频中重建高真实感的可渲染人体模型，本文提出了一种新颖的动态人体表示方法。本方法定义了一组可学习的隐变量，并将这组隐变量固定在一个可变形人体模型的网格顶点上，从而可以通过人体姿态变换这组隐变量的空间位置。为了表示不同视频帧下的数字人体，本方法将隐变量变换到对应的人体姿态下，然后使用一个三维卷积神经网络处理结构化隐变量，将离散的隐变量弥散为一个特征体，然后通过三线性插值得到任意一个三维点的特征向量。最后，本方法使用一个 MLP 网络从三维点的特征向量中回归得到三维点的体素密度和颜色。通过从一组相同的隐变量中得到不同视频帧下的三维人体，本方法在优化过程中融合了不同时刻的观测信息，从而获得足够的观测以解决稀疏视角重建的病态问题。为了验证上述提出的动态人体表示，本文采集了一个多视角视频数据集。在该数据集上的实验结果证明了本方法的新视角合成效果远远超出了之前的方法。该动态人体表示也启发了诸多的后续工作，比如 Neural Human Performer[67]、GP-NeRF[243]、Relighting4D[91]。

（2）为了实现可通过人体姿态进行驱动的神经数字人，本文利用骨骼蒙皮驱动算法构建了一种新颖的空间变形场。本方法提出了神经蒙皮权重场，使用一个 MLP 网络预测任意三维点的蒙皮权重向量，然后将蒙皮权重与从图片中估计得到的三维人体姿态参

数进行线性加权求和，得到一个三维点从世界坐标系到标准坐标系的变换矩阵。为了提升优化过程的稳定性，本方法充分利用了三维人体姿态参数的先验。具体而言，本方法利用人体姿态得到相应的参数化模型，然后得到目标姿态下的初始蒙皮权重场，从而给神经蒙皮权重场提供了优化起始点。此外，为了获得新姿态下的数字人，本方法利用循环一致性优化得到标准空间坐标系下的神经蒙皮权重场，然后再利用标准一致性获得新的人体姿态下的蒙皮权重场，最后通过骨骼蒙皮驱动算法产生从新人体姿态坐标系到标准坐标系的变形场。本方法在 Human3.6M[23] 和 ZJU-MoCap[250] 两个真实数据集上进行了充分验证。实验结果表明，本方法不仅能实现高真实感的新视角合成，还在新姿态合成方面远远超过了之前的方法。该工作为数字人建模提供了新的研究思路，启发了诸多的研究工作，比如 HumanNeRF[90]、ARAH[213]、NeuMan[254]。

（3）为了解决神经人体表示的几何重建质量较差的问题，本文提出了一种新颖的动态人体几何表示方法。针对之前方法中人体几何欠约束的问题，本方法使用符号距离场建模标准坐标系下的三维几何，利用其需要满足程函方程的特性对优化过程施加有效的正则化，从而保证重建得到的人体几何的平滑性。为了从 RGB 视频中优化人体几何，本方法借助变形场将世界坐标系与标准坐标系建立关联，并使用了基于符号距离场的体积渲染合成人体的图片。具体而言，本方法首先将符号距离场转化为体素密度，然后再用体积渲染将体素密度场投影为图片。为了建模任意人体姿态下的变形场，本方法将人体运动解耦为铰链式运动和非刚体运动，其中铰链式运动由参数化人体模型进行建模，而非刚体由一个基于隐式神经表示的位移场进行表示。该位移场将目标人体姿态参数作为神经网络的输入，然后预测任意三维点在标准坐标系下的位移，从而有效地建模了动态人体模型中非刚性的运动。本文在多个数据集上验证了上述提出的方法的有效性。实验结果表明，本文所提出的动态人体几何模型在人体几何重建方面大幅度地超过了之前的方法，并进一步提升了新姿态合成的渲染质量。

（4）为了实现动态场景的实时高真实感渲染，本文探索了一种新颖的体积视频表示，称为动态 MLP 图。该 MLP 图是一个二维网格，其中每个网格像素存放了一个 MLP 网络的参数向量。基于 MLP 图，本方法实现了通过一组小型 MLP 网络表示三维场景，从而降低 MLP 网络的推理成本，提升了渲染速度。为了高效地记录动态 MLP 图，本方法构造了一个超网络，使用二维卷积神经网络预测 MLP 图，从而降低了场景模型的存储成本。针对于高真实感渲染的问题，本方法使用了一个多层哈希表存储特征向量，用于

赋予输入的三维点一个高维的特征向量，从而提升本模型的表达能力。为了降低多层哈希表的存储成本，本方法使用 Tri-plane 的形式表示多层哈希表，在 XY、XZ、YZ 平面分别定义一个小型的多层哈希表，再通过加和得到任意三维点的特征向量。为了进一步提升渲染速度，本方法预先计算了空间三维点的体素密度，并记录在一个三维体素网格中。基于这个体素密度网格，本方法在渲染时直接跳过没有场景内容的三维区域，从而降低了计算成本。本文在 NHR 和 ZJU-MoCap 数据集上进行了充分的实验。实验结果表明，相比于之前的方法，动态 MLP 图在渲染速度和渲染质量上表现出最好的效果。

7.2 未来发展方向的展望

本文研究了动态三维人体的建模与渲染技术，分别从降低采集设备复杂度、实现可驱动性、提升几何精度、加速渲染过程等方面推动了数字人领域的进步，并为后续数字人的研究提供了新的思路。尽管本文实现了从 RGB 视频中建模高真实感的数字人，但这只是数字人领域的一个方向。为了进一步扩大数字虚拟人在内容创作、元宇宙、影视等方面的应用，数字人领域有着其他几个亟需探索的方向：

（1）面向数字人的生成式模型（Generative model）[269-272]。本文从图片中重建数字人体，高度还原了所观测到的人体。但是在实际应用中，一些用户出于隐私的考虑，往往希望以虚构的人体形象出现在网络平台。生成式模型通过学习目标数据的分布，再从数据分布中采样以生成虚构的数据，所以可以满足用户的这一需求。此外，生成式模型只需从分布中采样即可输出数字人，具有大规模创作数字人的能力。因此，数字人生成式模型具有很强的研究与应用价值。

（2）数字人的高自由度编辑。在虚拟世界中，用户大多会希望能装扮自身的虚拟形象以满足社交娱乐等方面的需求，比如更换数字人的衣服、给数字人化妆、改变数字人的发型。除此之外，内容创作领域的应用也会希望高自由度地操作数字人以制作特定的数字内容。因此，数字人编辑具有很高的应用价值。如何利用深度学习模型实现高自由度且精准的编辑能力，并支持多模态的用户输入，是未来重要的一个研究方向。

（3）具备智能系统的数字人。除了受真人驱动的数字人，由智能系统驱动的数字虚拟人也有着很强的应用价值。此类数字人能自主地感知外部环境，做出决策，并执行相应的动作，从而与真实世界的人类进行交互。例如人们与数字人进行对话时，数字人可

以通过观察人们的情绪做出合适的反应。具备智能系统的数字人对于虚拟伴侣、智能助手等应用至关重要。如何构建具有感知能力和决策能力的数字人是计算机视觉与人工智能领域非常重要的一个研究问题。

参考文献

[1] GUO K, LINCOLN P, DAVIDSON P, et al. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting[J]. ACM Transactions on Graphics, 2019, 38(6): 217:1-217:19.

[2] LAWRENCE J, GOLDMAN D, ACHAR S, et al. Project Starline: A High-Fidelity Telepresence System[J]. ACM Transactions on Graphics, 2021, 40(6): 242:1-242:16.

[3] ZHAO F, JIANG Y, YAO K, et al. Human Performance Modeling and Rendering via Neural Animated Mesh[J]. ACM Transactions on Graphics, 2022, 41(6): 235:1-235:17.

[4] NEWCOMBE R A, FOX D, SEITZ S M. Dynamicfusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015.

[5] INNMANN M, ZOLLHOFER M, NIESSNER M, et al. VolumeDeform: Real-Time Volumetric Non-rigid Reconstruction[C]//Proceedings of the European Conference on Computer Vision. 2016.

[6] YU T, GUO K, XU F, et al. BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2017.

[7] DOU M, KHAMIS S, DEGTYAREV Y, et al. Fusion4d: Real-Time Performance Capture of Challenging Scenes[J]. ACM Transactions on Graphics, 2016, 35(4): 114:1-114:13.

[8] DEBEVEC P, HAWKINS T, TCHOU C, et al. Acquiring The Reflectance Field of a Human Face[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 2000.

[9] COLLET A, CHUANG M, SWEENEY P, et al. High-Quality Streamable Free-Viewpoint Video[J]. ACM Transactions on Graphics, 2015, 34(4): 69:1-69:13.

[10] GORTLER S J, GRZESZCZUK R, SZELISKI R, et al. The Lumigraph[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1996.

[11] DAVIS A, LEVOY M, DURAND F. Unstructured Light Fields[C]//Eurographics. 2012.

[12] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy Networks: Learning 3D Reconstruction in Function Space[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[13] SITZMANN V, ZOLLHOFER M, WETZSTEIN G. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations[C]//Advances in Neural Information Processing Systems. 2019.

[14] PARK J J, FLORENCE P, STRAUB J, et al. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[15] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[C]//Proceedings of the European Conference on Computer Vision. 2020.

[16] LIU S, ZHANG Y, PENG S, et al. DIST: Rendering Deep Implicit Signed Distance Function with Differentiable Sphere Tracing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[17] YARIV L, GU J, KASTEN Y, et al. Volume Rendering of Neural Implicit Surfaces[C]//Advances in Neural Information Processing Systems. 2021.

[18] WANG P, LIU L, LIU Y, et al. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction[C]//Advances in Neural Information Processing Systems. 2021.

[19] HA D, DAI A, LE Q V. Hypernetworks[J]. ArXiv preprint arXiv:1609.09106, 2016.

[20] ALLDIECK T, MAGNOR M, BHATNAGAR B L, et al. Learning to Reconstruct People in Clothing from a Single RGB Camera[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[21] LEWIS J P, CORDNER M, FONG N. Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 2000.

[22] JOO H, SIMON T, SHEIKH Y. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[23] IONESCU C, PAPAVA D, OLARU V, et al. HUMAN3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325-1339.
[24] HART J C. Sphere Tracing: A Geometric Method for the Antialiased Ray Tracing of Implicit Surfaces[J]. The Visual Computer, 1996, 12(10): 527-545.
[25] WU M, WANG Y, HU Q, et al. Multi-View Neural Human Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[26] LI T, SLAVCHEVA M, ZOLLHOEFER M, et al. Neural 3D Video Synthesis From Multi-View Video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[27] YU T, ZHENG Z, GUO K, et al. DoubleFusion: Real-Time Capture of Human Performances with Inner Body Shapes from a Single Depth Sensor[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[28] SU Z, XU L, ZHENG Z, et al. RobustFusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera[C]//Proceedings of the European Conference on Computer Vision. 2020.
[29] LI Z, YU T, ZHENG Z, et al. POSEFusion: Pose-guided Selective Fusion for Single-view Human Volumetric Capture[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[30] YU T, ZHENG Z, GUO K, et al. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[31] BOGO F, KANAZAWA A, LASSNER C, et al. Keep it SMPL: Automatic Estimation of 3D human Pose and Shape from A Single Image[C]//Proceedings of the European Conference on Computer Vision. 2016.
[32] KANAZAWA A, BLACK M J, JACOBS D W, et al. End-to-End Recovery of Human Shape and Pose[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[33] KOLOTOUROS N, PAVLAKOS G, DANIILIDIS K. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[34] ALLDIECK T, PONS-MOLL G, THEOBALT C, et al. Tex2Shape: Detailed Full Human Body Geometry from a Single Image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[35] HABERMANN M, XU W, ZOLLHOFER M, et al. DeepCap: Monocular Human Performance Capture Using Weak Supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[36] LIN K, WANG L, LIU Z. End-to-End Human Pose and Mesh Reconstruction with Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[37] KOLOTOUROS N, PAVLAKOS G, BLACK M J, et al. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[38] SUN Y, BAO Q, LIU W, et al. Monocular, One-Stage, Regression of Multiple 3D People[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[39] LI J, XU C, CHEN Z, et al. Hybrik: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[40] ZHANG H, TIAN Y, ZHOU X, et al. Pymaf: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[41] DONG J, JIANG W, HUANG Q, et al. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[42] ZHANG Y, AN L, YU T, et al. 4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[43] ZHOU Z, SHUAI Q, WANG Y, et al. QuickPose: Real-time Multi-view Multi-person Pose Estimation in Crowded Scenes[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference. 2022.
[44] KANAZAWA A, ZHANG J Y, FELSEN P, et al. Learning 3D Human Dynamics From Video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[45] KOCABAS M, ATHANASIOU N, BLACK M J. VIBE: Video Inference for Human Body Pose and Shape Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[46] REMPE D, BIRDAL T, HERTZMANN A, et al. Humor: 3D Human Motion Model for Robust Pose Estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[47] XU W, CHATTERJEE A, ZOLLHOFER M, et al. MonoPerfCap: Human Performance Capture from Monocular Video[J]. ACM Transactions on Graphics, 2018, 37(2): 27.
[48] HABERMANN M, XU W, ZOLLHOEFER M, et al. Livecap: Real-Time Human Performance Capture from Monocular Video[J]. ACM Transactions on Graphics, 2019, 38(2): 1-17.
[49] MEHTA D, SOTNYCHENKO O, MUELLER F, et al. XNect: Real-Time Multi-Person 3D Motion Capture with a Single Rgb Camera[J]. Acm Transactions on Graphics, 2020, 39(4): 82.
[50] YUAN Y, WEI S E, SIMON T, et al. Simpoe: Simulated Character Control for 3D Human Pose Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[51] YUAN Y, IQBAL U, MOLCHANOV P, et al. GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[52] VAROL G, CEYLAN D, RUSSELL B, et al. Bodynet: Volumetric Inference of 3D Human Body Shapes[C]//Proceedings of the European Conference on Computer Vision. 2018.
[53] ZHENG Z, YU T, WEI Y, et al. DeepHuman: 3D Human Reconstruction From a Single Image[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[54] GILBERT A, VOLINO M, COLLOMOSSE J, et al. Volumetric Performance Capture from Minimal Camera Viewpoints[C]//Proceedings of the European Conference on Computer Vision. 2018.
[55] CALISKAN A, MUSTAFA A, HILTON A. Temporal Consistency Loss for High Resolution Textured and Clothed 3D Human Reconstruction from Monocular Video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[56] SAITO S, HUANG Z, NATSUME R, et al. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[57] SAITO S, SIMON T, SARAGIH J, et al. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[58] HE T, COLLOMOSSE J, JIN H, et al. Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-View Human Reconstruction[C]//Advances in Neural Information Processing Systems. 2020.
[59] LI R, XIU Y, SAITO S, et al. Monocular Real-Time Volumetric Performance Capture[C]//Proceedings of the European Conference on Computer Vision. 2020.
[60] HUANG Z, XU Y, LASSNER C, et al. ARCH: Animatable Reconstruction of Clothed Humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[61] ZHENG Z, YU T, LIU Y, et al. Pamir: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6): 3170-3184.

[62] HE T, XU Y, SAITO S, et al. Arch++: Animation-Ready Clothed Human Reconstruction Revisited[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[63] XIU Y, YANG J, TZIONAS D, et al. ICON: Implicit Clothed Humans Obtained from Normals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[64] HUANG Z, LI T, CHEN W, et al. Deep Volumetric Video From Very Sparse Multi-View Performance Capture[C]//Proceedings of the European Conference on Computer Vision. 2018.
[65] HONG Y, ZHANG J, JIANG B, et al. Stereopifu: Depth Aware Clothed Human Digitization via Stereo Vision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[66] SHAO R, ZHANG H, ZHANG H, et al. Doublefield: Bridging the Neural Surface and Radiance Fields for High-Fidelity Human Reconstruction and Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[67] KWON Y, KIM D, CEYLAN D, et al. Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering[C]//Advances in Neural Information Processing Systems. 2021.
[68] ZHENG Y, SHAO R, ZHANG Y, et al. Deepmulticap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[69] BARRON J T, MILDENHALL B, TANCIK M, et al. Mip-NeRF: A Multiscale Representation for Anti-aliasing Neural Radiance Fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[70] BARRON J T, MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: Unbounded Anti-aliased Neural Radiance Fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[71] GARBIN S J, KOWALSKI M, JOHNSON M, et al. Fastnerf: High-Fidelity Neural Rendering at 200FPS[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[72] LIU L, GU J, LIN K Z, et al. Neural Sparse Voxel Fields[C]//Advances in Neural Information Processing Systems. 2020.
[73] YU A, LI R, TANCIK M, et al. Plenoctrees for Real-Time Rendering of Neural Radiance Fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[74] REISER C, PENG S, LIAO Y, et al. KiloNeRF: Speeding Up Neural Radiance Fields with Thousands of Tiny Mlps[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[75] YU A, FRIDOVICH-KEIL S, TANCIK M, et al. Plenoxels: Radiance Fields without Neural Networks[C]//Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision. 2021.
[76] SUN C, SUN M, CHEN H T. Direct Voxel Grid Optimization: Super-Fast Convergence for Radiance Fields Reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[77] MULLER T, EVANS A, SCHIED C, et al. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding[J]. ACM Transactions on Graphics, 2022, 41(4): 102:1-102:15.
[78] CHEN A, XU Z, GEIGER A, et al. TensoRF: Tensorial Radiance Fields[C]//Proceedings of the European Conference on Computer Vision. 2022.
[79] ZHANG X, SRINIVASAN P P, DENG B, et al. Nerfactor: Neural Factorization of Shape and Reflectance under an Unknown Illumination[J]. ACM Transactions on Graphics, 2021, 40(6): 237:1-237:18.
[80] ZHANG Y, SUN J, HE X, et al. Modeling Indirect Illumination for Inverse Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[81] LOMBARDI S, SIMON T, SARAGIH J, et al. Neural Volumes: Learning Dynamic Renderable Volumes from Images[J]. ACM Transactions on Graphics, 2019, 38(4): 65:1-65:14.
[82] PARK K, SINHA U, BARRON J T, et al. Nerfies: Deformable Neural Radiance Fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[83] PENG S, ZHANG Y, XU Y, et al. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[84] PUMAROLA A, CORONA E, PONS-MOLL G, et al. D-NeRF: Neural Radiance Fields for Dynamic Scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[85] LIU L, HABERMANN M, RUDNEV V, et al. Neural Actor: Neural Free-View Synthesis of Human Actors with Pose Control[J]. ACM Transactions on Graphics, 2021, 40(6): 219:1-219:16.
[86] PENG S, DONG J, WANG Q, et al. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[87] LOMBARDI S, SIMON T, SCHWARTZ G, et al. Mixture of Volumetric Primitives for Efficient Neural Rendering[J]. ACM Transactions on Graphics, 2021, 40(4): 59:1-59:13.
[88] WANG L, ZHANG J, LIU X, et al. Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-time[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[89] SHUAI Q, GENG C, FANG Q, et al. Novel View Synthesis of Human Interactions from Sparse Multi-View Videos[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference. 2022.
[90] WENG C Y, CURLESS B, SRINIVASAN P P, et al. HumanNeRF: Free-Viewpoint Rendering of Moving People from Monocular Video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[91] CHEN Z, LIU Z. Relighting4d: Neural Relightable Human from Videos[C]//Proceedings of the European Conference on Computer Vision. 2022.
[92] FANG J, YI T, WANG X, et al. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels[C]//SIGGRAPH Asia 2022 Conference Papers. 2022.
[93] JIANG T, CHEN X, SONG J, et al. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds[J]. ArXiv preprint arXiv:2212.10550, 2022.
[94] GAO X, YANG J, KIM J, et al. MPS-NeRF: Generalizable 3D Human Rendering from Multiview Images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[95] ZHENG E, DUNN E, JOJIC V, et al. Patchmatch Based Joint View Selection and Depthmap Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2014.
[96] SCHONBERGER J L, FRAHM J M. Structure-from-Motion Revisited[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016.
[97] SCHONBERGER J L, ZHENG E, FRAHM J M, et al. Pixelwise View Selection for Unstructured Multi-View Stereo[C]//Proceedings of the European Conference on Computer Vision. 2016.
[98] HECKBERT P S. Survey of Texture Mapping[J]. IEEE computer graphics and applications, 1986, 6(11): 56-67.
[99] HECKBERT P S. Fundamentals of Texture Mapping and Image Warping[J]., 1989.
[100] KAZHDAN M, BOLITHO M, HOPPE H. Poisson Surface Reconstruction[C]//Proceedings of the Eurographics symposium on Geometry processing. 2006.
[101] ZHENG Z, YU T, LI H, et al. Hybridfusion: Real-Time Performance Capture Using a Single Depth Sensor and Sparse IMUs[C]//Proceedings of the European Conference on Computer Vision. 2018.
[102] SUMNER R W, SCHMID J, PAULY M. Embedded Deformation for Shape Manipulation[J]. ACM Transactions on Graphics, 2007, 26(3): 80.
[103] CURLESS B, LEVOY M. A Volumetric Method for Building Complex Models from Range Images[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1996.
[104] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A Skinned Multi-person Linear Model[J]. ACM Transactions on Graphics, 2015, 34(6): 248:1-248:16.
[105] DOU M, DAVIDSON P, FANELLO S R, et al. Motion2fusion: Real-Time Volumetric Performance Capture[J]. ACM Transactions on Graphics, 2017, 36(6): 246:1-246:16.

[106] SU Z, XU L, ZHENG Z, et al. RobustFusion: Human Volumetric Capture with Data-Driven Visual Cues Using a RGBD Camera[C]//Proceedings of the European Conference on Computer Vision. 2020.
[107] PAVLAKOS G, ZHU L, ZHOU X, et al. Learning to Estimate 3D Human Pose and Shape from a Single Color Image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[108] PAVLAKOS G, CHOUTAS V, GHORBANI N, et al. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[109] XU H, BAZAVAN E G, ZANFIR A, et al. Ghum & Ghuml: Generative 3D Human Shape and Articulated Pose Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[110] ANGUELOV D, SRINIVASAN P, KOLLER D, et al. Scape: Shape Completion and Animation of People[J]. ACM Transactions on Graphics, 2005, 24(3): 408-416.
[111] PONS-MOLL G, ROMERO J, MAHMOOD N, et al. Dyna: A Model of Dynamic Human Shape in Motion[J]. ACM Transactions on Graphics, 2015, 34(4): 120:1-120:14.
[112] OSMAN A A, BOLKART T, BLACK M J. STAR: Sparse Trained Articulated Human Body Regressor[C]//Proceedings of the European Conference on Computer Vision. 2020.
[113] WANG H, GULER R A, KOKKINOS I, et al. BLSM: A Bone-Level Skinned Model of the Human Mesh[C]//Proceedings of the European Conference on Computer Vision. 2020.
[114] SANTESTEBAN I, GARCES E, OTADUY M A, et al. SoftSMPL: Data-driven Modeling of Nonlinear Soft-tissue Dynamics for Parametric Humans[C]//Computer Graphics Forum: vol. 39: 2. 2020: 65-75.
[115] DENG B, LEWIS J, JERUZALSKI T, et al. NASA: Neural Articulated Shape Approximation[C]//Proceedings of the European Conference on Computer Vision. 2020.
[116] ALLDIECK T, XU H, SMINCHISESCU C. Imghum: Implicit Generative Models of 3D Human Shape and Articulated Pose[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[117] ZANFIR M, ALLDIECK T, SMINCHISESCU C. PhoMoH: Implicit Photorealistic 3D Models of Human Heads[J]. ArXiv preprint arXiv:2212.07275, 2022.
[118] HONG Y, PENG B, XIAO H, et al. Headnerf: A Real-Time Nerf-Based Parametric Head Model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[119] ZHUANG Y, ZHU H, SUN X, et al. Mofanerf: Morphable Facial Neural Radiance Field[C]//Proceedings of the European Conference on Computer Vision. 2022.
[120] TIWARI G, ANTIC D, LENSSEN J E, et al. Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields[C]//Proceedings of the European Conference on Computer Vision. 2022.
[121] DONG J, SHUAI Q, ZHANG Y, et al. Motion Capture from Internet Videos[C]//Proceedings of the European Conference on Computer Vision. 2020.
[122] TIAN Y, ZHANG H, LIU Y, et al. Recovering 3D Human Mesh from Monocular Images: A Survey[J]. ArXiv preprint arXiv:2203.01923, 2022.
[123] JIANG W, KOLOTOUROS N, PAVLAKOS G, et al. Coherent Reconstruction of Multiple Humans from a Single Image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[124] KIPF T N, WELLING M. Semi-supervised Classification with Graph Convolutional Networks[J]. ArXiv preprint arXiv:1609.02907, 2016.
[125] LITANY O, BRONSTEIN A, BRONSTEIN M, et al. Deformable Shape Completion with Graph Convolutional Autoencoders[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[126] TU H, WANG C, ZENG W. Voxelpose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment[C]//Proceedings of the European Conference on Computer Vision. 2020.
[127] ZHANG J, CAI Y, YAN S, et al. Direct Multi-View Multi-Person 3D Pose Estimation[C]//Advances in Neural Information Processing Systems. 2021.

[128] WU S, JIN S, LIU W, et al. Graph-based 3D Multi-Person Pose Estimation Using Multi-View Images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[129] YE H, ZHU W, WANG C, et al. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection[C]//Proceedings of the European Conference on Computer Vision. 2022.
[130] LIN J, LEE G H. Multi-view Multi-Person 3D Pose Estimation with Plane Sweep Stereo[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[131] ALLDIECK T, MAGNOR M, XU W, et al. Video Based Reconstruction of 3D People Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[132] ZHU L, REMATAS K, CURLESS B, et al. Reconstructing NBA Players[C]//Proceedings of the European Conference on Computer Vision. 2020.
[133] ALLDIECK T, MAGNOR M, XU W, et al. Detailed Human Avatars from Monocular Video[C]//2018 International Conference on 3D Vision (3DV). 2018.
[134] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 43(1): 172-186.
[135] SUN K, XIAO B, LIU D, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[136] DONG Z, XU K, DUAN Z, et al. Geometry-aware Two-scale PIFu Representation for Human Reconstruction[C]//Advances in Neural Information Processing Systems. 2022.
[137] CHAN K Y, LIN G, ZHAO H, et al. Integratedpifu: Integrated Pixel Aligned Implicit Function for Single-View Human Reconstruction[C]//Proceedings of the European Conference on Computer Vision. 2022.
[138] ALLDIECK T, ZANFIR M, SMINCHISESCU C. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[139] CORONA E, ZANFIR M, ALLDIECK T, et al. Structured 3D Features for Reconstructing Relightable and Animatable Avatars[J]. ArXiv preprint arXiv:2212.06820, 2022.
[140] CHIBANE J, ALLDIECK T, PONS-MOLL G. Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[141] BHATNAGAR B L, SMINCHISESCU C, THEOBALT C, et al. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction[C]//Proceedings of the European Conference on Computer Vision. 2020.
[142] YANG Z, WANG S, MANIVASAGAM S, et al. S3: Neural Shape, Skeleton, and Skinning Fields for 3D Human Modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[143] SAITO S, YANG J, MA Q, et al. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[144] QI C R, YI L, SU H, et al. Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]//Advances in Neural Information Processing Systems. 2017.
[145] SUO X, JIANG Y, LIN P, et al. Neuralhumanfvv: Real-Time Neural Volumetric Human Performance Rendering Using Rgb Cameras[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[146] FENG Q, LIU Y, LAI Y K, et al. FOF: Learning Fourier Occupancy Field for Monocular Real-Time Human Reconstruction[C]//Advances in Neural Information Processing Systems. 2022.
[147] YAO G, WU H, YUAN Y, et al. DD-NeRF: Double-Diffusion Neural Radiance Field as a Generalizable Implicit Body Representation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2021.
[148] NIMIER-DAVID M, VICINI D, ZELTNER T, et al. Mitsuba 2: A Retargetable Forward and Inverse Renderer[J]. ACM Transactions on Graphics, 2019, 38(6): 203:1-203:17.

[149] KATO H, BEKER D, MORARIU M, et al. Differentiable Rendering: A Survey[J]. ArXiv preprint arXiv:2006.12057, 2020.
[150] LIOR Y, YONI K, DROR M, et al. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance[C]//Advances in Neural Information Processing Systems. 2020.
[151] TEWARI A, FRIED O, THIES J, et al. State of the Art on Neural Rendering[C]//Computer Graphics Forum: vol. 39: 2. 2020: 701-727.
[152] TEWARI A, THIES J, MILDENHALL B, et al. Advances in Neural Rendering[C]//Computer Graphics Forum: vol. 41: 2. 2022: 703-735.
[153] LIU S, LI T, CHEN W, et al. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[154] YIFAN W, SERENA F, WU S, et al. Differentiable Surface Splatting for Point-Based Geometry Processing[J]. ACM Transactions on Graphics, 2019, 38(6): 230:1-230:14.
[155] RAVI N, REIZENSTEIN J, NOVOTNY D, et al. Accelerating 3D Deep Learning with PyTorch3D[J]. ArXiv:2007.08501, 2020.
[156] LOPER M M, BLACK M J. OpenDR: An Approximate Differentiable Renderer[C]//Proceedings of the European Conference on Computer Vision. 2014.
[157] KATO H, USHIKU Y, HARADA T. Neural 3D Mesh Renderer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[158] HU Y, LI T M, ANDERSON L, et al. Taichi: A Language for High-Performance Computation on Spatially Sparse Data Structures[J]. ACM Transactions on Graphics, 2019, 38(6): 201:1-201:16.
[159] THIES J, ZOLLHOFER M, NIESSNER M. Deferred Neural Rendering: Image Synthesis Using Neural Textures[J]. ACM Transactions on Graphics, 2019, 38(4): 66:1-66:12.
[160] RAJ A, TANKE J, HAYS J, et al. ANR: Articulated Neural Rendering for Virtual Avatars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[161] CHEN Z, FUNKHOUSER T, HEDMAN P, et al. Mobilenerf: Exploiting the Polygon Rasterization Pipeline for Efficient Neural Field Rendering on Mobile Architectures[C]//. 2022.
[162] LASSNER C, ZOLLHOFER M. Pulsar: Efficient Sphere-Based Neural Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[163] ALIEV K A, SEVASTOPOLSKY A, KOLOS M, et al. Neural Point-Based Graphics[C]//Proceedings of the European Conference on Computer Vision. 2020.
[164] NIEMEYER M, MESCHEDER L, OECHSLE M, et al. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[165] KAJIYA J T, VON HERZEN B P. Ray Tracing Volume Densities[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1984.
[166] BI S, XU Z, SRINIVASAN P, et al. Neural Reflectance Fields for Appearance Acquisition[J]. ArXiv preprint arXiv:2008.03824, 2020.
[167] BI S, XU Z, SUNKAVALLI K, et al. Deep Reflectance Volumes: Relightable Reconstructions from Multi-View Photometric Images[C]//. 2020.
[168] REBAIN D, JIANG W, YAZDANI S, et al. Derf: Decomposed Radiance Fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[169] HEDMAN P, SRINIVASAN P P, MILDENHALL B, et al. Baking Neural Radiance Fields for Real-Time View Synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[170] ZHANG J, HUANG J, CAI B, et al. Digging into Radiance Grid for Real-Time View Synthesis with Detail Preservation[C]//Proceedings of the European Conference on Computer Vision. 2022.
[171] HU T, LIU S, CHEN Y, et al. EfficientNeRF: Efficient Neural Radiance Fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[172] NEFF T, STADLBAUER P, PARGER M, et al. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks[C]//Computer Graphics Forum: vol. 40: 4. 2021: 45-59.
[173] LIN H, PENG S, XU Z, et al. Efficient Neural Radiance Fields for Interactive Free-viewpoint Video[C]//SIGGRAPH Asia 2022 Conference Papers. 2022.

[174] KURZ A, NEFF T, LV Z, et al. AdaNeRF: Adaptive Sampling for Real-Time Rendering of Neural Radiance Fields[C]//Proceedings of the European Conference on Computer Vision. 2022.
[175] ARANDJELOVIC R, ZISSERMAN A. NeRF in Detail: Learning to Sample for View Synthesis[J]. ArXiv preprint arXiv:2106.05264, 2021.
[176] FANG J, XIE L, WANG X, et al. Neusample: Neural Sample Field for Efficient View Synthesis[J]. ArXiv preprint arXiv:2111.15552, 2021.
[177] DADON D, FRIED O, HEL-OR Y. DDNeRF: Depth Distribution Neural Radiance Fields[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[178] LINDELL D B, MARTEL J N, WETZSTEIN G. Autoint: Automatic Integration for Fast Neural Volume Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[179] WU L, LEE J Y, BHATTAD A, et al. Diver: Real-Time and Accurate Neural Radiance Fields with Deterministic Integration for Volume Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[180] SITZMANN V, REZCHIKOV S, FREEMAN B, et al. Light Field Networks: Neural Scene Representations with Single-Evaluation Rendering[C]//Advances in Neural Information Processing Systems. 2021.
[181] WANG H, REN J, HUANG Z, et al. R2l: Distilling Neural Radiance Field to Neural Light Field for Efficient Novel View Synthesis[C]//Proceedings of the European Conference on Computer Vision. 2022.
[182] TANCIK M, MILDENHALL B, WANG T, et al. Learned Initializations for Optimizing Coordinate-Based Neural Representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[183] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-Learning in Neural Networks: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 5149-5169.
[184] WANG S, MIHAJLOVIC M, MA Q, et al. MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images[C]//Advances in Neural Information Processing Systems. 2021.
[185] CHEN Y, WANG X. Transformers as Meta-learners for Implicit Neural Representations[C]//Proceedings of the European Conference on Computer Vision. 2022.
[186] TIAN Z, SHEN C, CHEN H. Conditional Convolutions for Instance Segmentation[C]//Proceedings of the European Conference on Computer Vision. 2020.
[187] MAXIMOV M, LEAL-TAIXE L, FRITZ M, et al. Deep Appearance Maps[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.
[188] CHEN Y, DAI X, LIU M, et al. Dynamic Convolution: Attention over Convolution Kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[189] CHIBANE J, BANSAL A, LAZOVA V, et al. Stereo Radiance Fields (SRF): Learning View Synthesis for Sparse Views of Novel Scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[190] WANG Q, WANG Z, GENOVA K, et al. IBRNet: Learning Multi-View Image-Based Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[191] CHEN A, XU Z, ZHAO F, et al. Mvsnerf: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[192] LIU Y, PENG S, LIU L, et al. Neural Rays for Occlusion-Aware Image-Based Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[193] JOHARI M M, LEPOITTEVIN Y, FLEURET F. Geonerf: Generalizing NeRF with Geometry Priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[194] RAJ A, ZOLLHOFER M, SIMON T, et al. Pixel-Aligned Volumetric Avatars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[195] SRINIVASAN P P, DENG B, ZHANG X, et al. Nerv: Neural Reflectance and Visibility Fields for Relighting and View Synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[196] ZHANG K, LUAN F, LI Z, et al. IRON: Inverse Rendering by Optimizing Neural Sdfs and Materials from Photometric Imagess[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[197] ZHANG K, LUAN F, WANG Q, et al. PhySG: Inverse Rendering with Spherical Gaussians for Physics-Based Material Editing and Relighting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[198] PHARR M, JAKOB W, HUMPHREYS G. Physically Based Rendering: From Theory to Implementation[M]. 2016.
[199] YU A, YE V, TANCIK M, et al. PixelNeRF: Neural Radiance Fields from One or Few Images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[200] BURKOV E, RAKHIMOV R, SAFIN A, et al. Multi-NeuS: 3D Head Portraits from Single Image with Neural Implicit Functions[J]. ArXiv preprint arXiv:2209.04436, 2022.
[201] RAMON E, TRIGINER G, ESCUR J, et al. H3D-Net: Few-Shot High-Fidelity 3D Head Reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[202] CHAN E R, LIN C Z, CHAN M A, et al. Efficient Geometry-Aware 3D Generative Adversarial Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[203] SCHWARZ K, LIAO Y, NIEMEYER M, et al. Graf: Generative Radiance Fields for 3D-Aware Image Synthesis[C]//Advances in Neural Information Processing Systems. 2020.
[204] XU Y, PENG S, YANG C, et al. 3D-Aware Image Synthesis via Learning Structural and Textural Representations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[205] HONG F, CHEN Z, LAN Y, et al. EVA3D: Compositional 3D Human Generation from 2D Image Collections[J]. ArXiv preprint arXiv:2210.04888, 2022.
[206] KO J, CHO K, CHOI D, et al. 3D Gan Inversion with Pose Optimization[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[207] XIE J, OUYANG H, PIAO J, et al. High-fidelity 3D GAN Inversion by Pseudo-multi-view Optimization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[208] GRAHAM B, VAN DER MAATEN L. Submanifold Sparse Convolutional Networks[J]. ArXiv preprint arXiv:1706.01307, 2017.
[209] GRAHAM B, ENGELCKE M, VAN DER MAATEN L. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[210] PARK K, SINHA U, HEDMAN P, et al. Hypernerf: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields[J]. ACM Transactions on Graphics, 2021, 40(6): 238:1-238:12.
[211] GAO C, SARAF A, KOPF J, et al. Dynamic View Synthesis from Dynamic Monocular Video[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[212] XU H, ALLDIECK T, SMINCHISESCU C. H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion[C]//Advances in Neural Information Processing Systems. 2021.
[213] WANG S, SCHWARZ K, GEIGER A, et al. ARAH: Animatable Volume Rendering of Articulated Human SDFs[C]//Proceedings of the European Conference on Computer Vision. 2022.
[214] ZHENG Z, HUANG H, YU T, et al. Structured Local Radiance Fields for Human Avatar Modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[215] RAJ A, TANKE J, HAYS J, et al. ANR: Articulated Neural Rendering for Virtual Avatars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[216] PROKUDIN S, BLACK M J, ROMERO J. SMPLpix: Neural Avatars from 3D Human Models[C]//WCCV. 2021.
[217] ZHANG R, CHEN J. NDF: Neural Deformable Fields for Dynamic Human Modelling[C]//Proceedings of the European Conference on Computer Vision. 2022.

[218] LI R, TANKE J, VO M, et al. Tava: Template-Free Animatable Volumetric Actors[C]//Proceedings of the European Conference on Computer Vision. 2022.
[219] HABERMANN M, LIU L, XU W, et al. Real-time Deep Dynamic Characters[J]. ACM Transactions on Graphics, 2021, 40(4): 94:1-94:16.
[220] LI Y, HABERMANN M, THOMASZEWSKI B, et al. Deep Physics-Aware Inference of Cloth Deformation for Monocular Human Performance Capture[C]//2021 International Conference on 3D Vision. 2021.
[221] JIANG Y, HABERMANN M, GOLYANIK V, et al. Hifecap: Monocular High-Fidelity and Expressive Capture of Human Performances[C]//British Machine Vision Conference. 2022.
[222] REMELLI E, BAGAUTDINOV T, SAITO S, et al. Drivable Volumetric Avatars Using Texel-Aligned Features[C]//Special Interest Group on Computer Graphics and Interactive Techniques Conference. 2022.
[223] ZHANG J, LIU X, YE X, et al. Editable Free-Viewpoint Video Using a Layered Neural Representation[J]. ACM Transactions on Graphics, 2021, 40(4): 149:1-149:18.
[224] LIU J W, CAO Y P, MAO W, et al. Devrf: Fast Deformable Voxel Radiance Fields for Dynamic Scenes[C]//Advances in Neural Information Processing Systems. 2022.
[225] CAO C, SIMON T, KIM J K, et al. Authentic Volumetric Avatars from a Phone Scan[J]. ACM Transactions on Graphics, 2022, 41(4): 163:1-163:19.
[226] BAI J, HUANG L, GONG W, et al. Self-NeRF: A Self-Training Pipeline for Few-Shot Neural Radiance Fields[J]. ArXiv preprint arXiv:2303.05775, 2023.
[227] GENG C, PENG S, XU Z, et al. Learning Neural Volumetric Representations of Dynamic Humans in Minutes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[228] HEDMAN P, PHILIP J, PRICE T, et al. Deep Blending for Free-Viewpoint Image-Based Rendering[J]. ACM Transactions on Graphics, 2018, 37(6): 257.
[229] DEBEVEC P E, TAYLOR C J, MALIK J. Modeling and Rendering Architecture from Photographs: A Hybrid Geometry-And Image-Based Approach[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1996.
[230] FANG Q, SHUAI Q, DONG J, et al. Reconstructing 3D Human Pose by Watching Humans in the Mirror[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[231] LOEHLIN J C. Latent Variable Models[M]. 1987.
[232] GONG K, LIANG X, LI Y, et al. Instance-Level Human Parsing via Part Grouping Network[C]//Proceedings of the European Conference on Computer Vision. 2018.
[233] JIANG C, SUD A, MAKADIA A, et al. Local Implicit Grid Representations for 3D Scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[234] CHABRA R, LENSSEN J E, ILG E, et al. Deep Local Shapes: Learning Local SDF Priors for Detailed 3D Reconstruction[C]//Proceedings of the European Conference on Computer Vision. 2020.
[235] GENOVA K, COLE F, SUD A, et al. Local Deep Implicit Functions for 3D Shape[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[236] YAN Y, MAO Y, LI B. SECOND: Sparsely Embedded Convolutional Detection[J]. Sensors, 2018, 18(10): 3337.
[237] SHI S, GUO C, JIANG L, et al. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[238] PENG S, NIEMEYER M, MESCHEDER L, et al. Convolutional Occupancy Networks[C]//Proceedings of the European Conference on Computer Vision. 2020.
[239] RAHAMAN N, BARATIN A, ARPIT D, et al. On the Spectral Bias of Neural Networks[C]//Proceedings of the International Conference on Machine Learning. 2019.
[240] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]//International Conference on Learning Representations. 2015.
[241] LORENSEN W E, CLINE H E. Marching Cubes: A High Resolution 3D Surface Construction Algorithm[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1987.

[242] QI C R, SU H, MO K, et al. Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[243] CHEN M, ZHANG J, XU X, et al. Geometry-guided Progressive Nerf for Generalizable and Efficient Neural Human Rendering[C]//Proceedings of the European Conference on Computer Vision. 2022.
[244] MIHAJLOVIC M, BANSAL A, ZOLLHOEFER M, et al. KeypointNeRF: Generalizing Image-Based Volumetric Avatars Using Relative Spatial Encoding of Keypoints[C]//Proceedings of the European Conference on Computer Vision. 2022.
[245] ZHI Y, QIAN S, YAN X, et al. Dual-Space NeRF: Learning Animatable Avatars and Scene Lighting in Separate Spaces[C]//International Conference on 3D Vision. 2022.
[246] XU T, FUJITA Y, MATSUMOTO E. Surface-Aligned Neural Radiance Fields for Controllable 3D Human Synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[247] PARK K, SINHA U, BARRON J T, et al. Nerfies: Deformable Neural Radiance Fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[248] PUMAROLA A, CORONA E, PONS-MOLL G, et al. D-NeRF: Neural Radiance Fields for Dynamic Scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[249] LI Z, NIKLAUS S, SNAVELY N, et al. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[250] PENG S, ZHANG Y, XU Y, et al. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[251] ROMERO J, TZIONAS D, BLACK M J. Embodied Hands: Modeling and Capturing Hands and Bodies Together[J]. ACM Transactions on Graphics, 2017, 36(6): 245:1-245:17.
[252] BHATNAGAR B L, SMINCHISESCU C, THEOBALT C, et al. LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration[C]//Advances in Neural Information Processing Systems. 2020.
[253] KAJIYA J T. The Rendering Equation[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1986.
[254] JIANG W, YI K M, SAMEI G, et al. Neuman: Neural Human Radiance Field from a Single Video[C]//Proceedings of the European Conference on Computer Vision. 2022.
[255] FYFFE G, WILSON C A, DEBEVEC P. Cosine Lobe Based Relighting from Gradient Illumination Photographs[C]//Conference for Visual Media Production. 2009.
[256] SEYB D, JACOBSON A, NOWROUZEZAHRAI D, et al. Non-linear Sphere Tracing for Rendering Deformed Signed Distance Fields[J]. ACM Transactions on Graphics, 2019, 38(6): 229:1-229:12.
[257] TIWARI G, SARAFIANOS N, TUNG T, et al. Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[258] GROPP A, YARIV L, HAIM N, et al. Implicit Geometric Regularization for Learning Shapes[C]//Proceedings of the International Conference on Machine Learning. 2020.
[259] SU S Y, YU F, ZOLLHOFER M, et al. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose[C]//Advances in Neural Information Processing Systems. 2021.
[260] LEVOY M, HANRAHAN P. Light Field Rendering[C]//Proceedings of the Conference on Computer Graphics and Interactive Techniques. 1996.
[261] BANSAL A, VO M, SHEIKH Y, et al. 4D Visualization of Dynamic Events from Unconstrained Multi-View Videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[262] YOON J S, KIM K, GALLO O, et al. Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[263] TURKI H, RAMANAN D, SATYANARAYANAN M. Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[264] TANCIK M, CASSER V, YAN X, et al. Block-NeRF: Scalable Large Scene Neural View Synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[265] MARTIN-BRUALLA R, RADWAN N, SAJJADI M S, et al. Nerf in the Wild: Neural Radiance Fields for Unconstrained Photo Collections[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[266] BOJANOWSKI P, JOULIN A, LOPEZ-PAZ D, et al. Optimizing the Latent Space of Generative Networks[J]. ArXiv preprint arXiv:1707.05776, 2017.
[267] WANG Z, BAGAUTDINOV T, LOMBARDI S, et al. Learning Compositional Radiance Fields of Dynamic Human Heads[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[268] ZHANG R, ISOLA P, EFROS A A, et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[269] KARRAS T, LAINE S, AILA T. A Style-Based Generator Architecture for Generative Adversarial Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[270] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and Improving the Image Quality of Stylegan[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[271] YANG L, ZHANG Z, SONG Y, et al. Diffusion Models: A Comprehensive Survey of Methods and Applications[J]. ArXiv preprint arXiv:2209.00796, 2022.
[272] BOND-TAYLOR S, LEACH A, LONG Y, et al. Deep Generative Modelling: A Comparative Review of Vaes, Gans, Normalizing Flows, Energy-Based and Autoregressive Models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 7327-7347.

攻读博士期间主要研究成果

[1] 第一作者 Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.（EI，CCF-A 类会议，CVPR 最佳论文候选，谷歌学术引用 268，GitHub 开源加星 782，本文第 3 章）

[2] 第一作者 Implicit Neural Representations with Structured Latent Codes for Human Body Modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2023.（SCI，CCF-A 类期刊，本文第 3 章）

[3] 第一作者 Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.（EI，CCF-A 类会议，谷歌学术引用 161，GitHub 开源加星 391，本文第 4 章）

[4] 第一作者 Representing Volumetric Videos as Dynamic MLP Maps. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.（EI，CCF-A 类会议，GitHub 开源加星 94，本文第 6 章）

[5] 第一作者 Animatable Implicit Neural Representations for Creating Realistic Avatars from Videos.（TPAMI 在投，本文第 5 章）

[6] 第一作者 PVNet: Pixel-wise Voting Network for 6DoF Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.（EI，CCF-A 类会议，CVPR 口头报告，谷歌学术引用 707，GitHub 开源加星 736）

[7] 第一作者 PVNet: Pixel-wise Voting Network for 6DoF Object Pose Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2020.（SCI，CCF-A 类期刊）

[8] 第一作者 Deep Snake for Real-Time Instance Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.（EI，CCF-A 类会议，CVPR 口头报告，谷歌学术引用 216，GitHub 开源加星 1100）

[9] 共同一作 Neural 3D Scene Reconstruction with the Manhattan-world Assumption. IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.（EI，CCF-A 类会议，CVPR 口头报告，谷歌学术引用 26，GitHub 开源加星 397）

[10] 共同一作 Learning Neural Volumetric Representations of Dynamic Humans in Minutes. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.（EI，CCF-A 类会议）

[11] 共同一作 Efficient Neural Radiance Fields for Interactive Free-viewpoint Video. ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia), 2022.（谷歌学术引用 13，GitHub 开源加星 283）

[12] 共同一作 Learning to Estimate Object Poses without Real Image Annotations. International Joint Conference on Artificial Intelligence (IJCAI), 2022.（EI，CCF-A 类会议）

[13] 第二作者 Painting 3D Nature in 2D: View Synthesis of Natural Scenes from a Single Semantic Mask. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.（EI，CCF-A 类会议）

[14] 第二作者 NerfCap: Human Performance Capture with Dynamic Neural Radiance Fields. IEEE Transactions on Visualization and Computer Graphics (TVCG), 2022.（SCI，CCF-A 类期刊）

[15] 第二作者 3D-aware Image Synthesis via Learning Structural and Textural Representations. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.（EI，CCF-A 类会议，谷歌学术引用 53，GitHub 开源加星 117）

[16] 第二作者 Neural Rays for Occlusion-aware Image-based Rendering. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.（EI，CCF-A 类会议，谷歌学术引用 44，GitHub 开源加星 330）

[17] 第二作者 Towards Efficient and Photorealistic 3D Human Reconstruction: A Brief Survey. Visual Informatics, 2022.