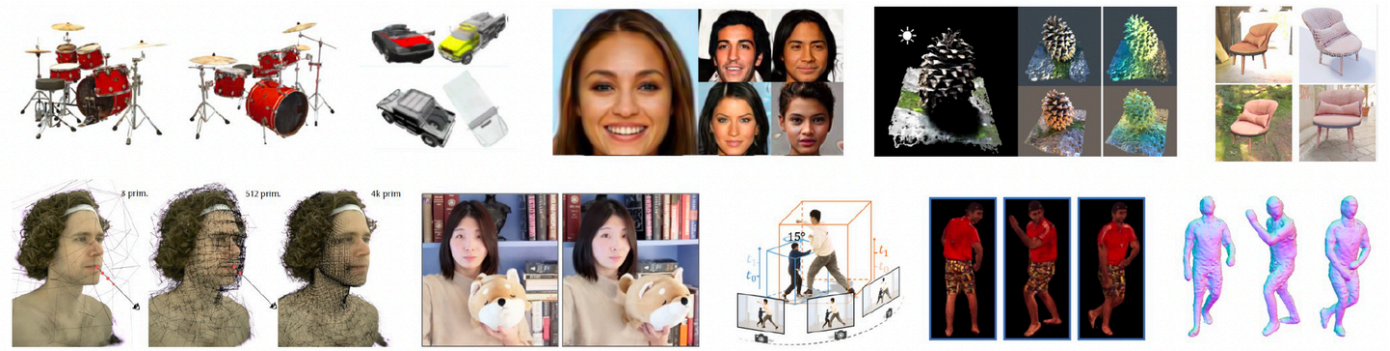
神经渲染的进展综述



EuroGraphics‘2022 综述论文“Advances in Neural Rendering“，2022年3月，作者来自MPI、谷歌研究、ETH、MIT、Reality Labs Research、慕尼黑工大和斯坦福大学。

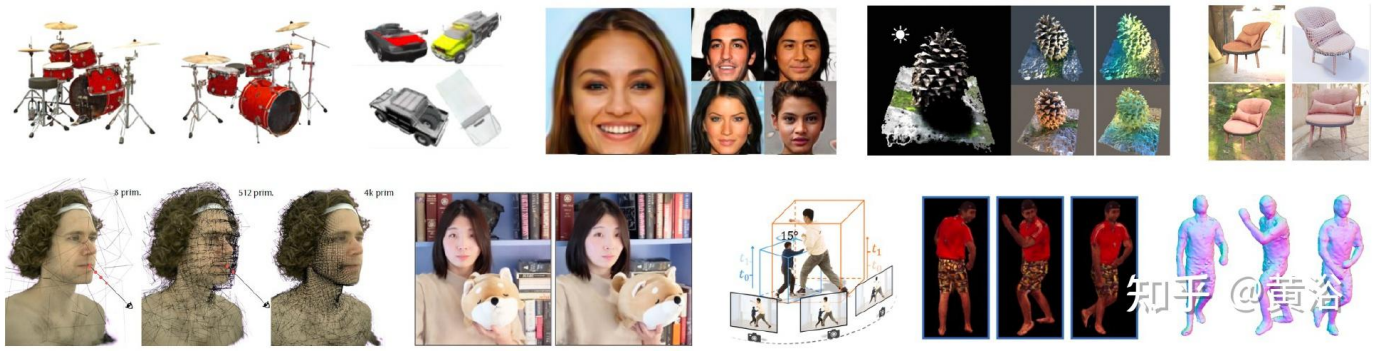
Advances in Neural Rendering

A. Tewari[1,6*] J. Thies[2*] B. Mildenhall[3*] P. Srinivasan[3*] E. Tretschk[1] W. Yifan[4,8] C. Lassner[5] V. Sitzmann[6] R. Martin-Brualla[3]
S. Lombardi[5] T. Simon[5] C. Theobalt[1] M. Nießner[7] J. T. Barron[3] G. Wetzstein[8] M. Zollhöfer[5] V. Golyanik[1]

[1]MPI for Informatics [2]MPI for Intelligent Systems [3]Google Research [4]ETH Zürich [5]Reality Labs Research
[6]MIT [7]Technical University of Munich [8]Stanford University *Equal contribution.

合成照片级逼真的图像和视频是计算机图形学的核心，也是几十年来研究的焦点。传统上，场景的合成图像是使用渲染算法（如光栅化或光线跟踪）生成的，这些算法将特别定义的几何和材质属性表示作为输入。总的来说，这些输入定义了实际场景和渲染的内容，称为**场景表示**（场景由一个或多个目标组成）。示例场景表示是具有伴随纹理的三角网格（例如，由艺术家创建）、点云（例如，来自深度传感器）、体网格（例如来自CT扫描）或隐式曲面函数（例如，截断符号距离场）。使用可微分渲染的损失从观测中重建这样的场景表示被称为逆图

形学或逆渲染。



神经渲染 是密切相关的，它结合了经典计算机图形学和机器学习的思想，创建了从真实世界观测合成图像的算法。神经渲染是朝着合成照片级逼真图像和视频内容的目标迈进的一步。近年来看到了该领域的巨大进步，展示了将可学习的组件注入渲染流水线的不同方法。

这篇关于神经渲染进展的最新报告侧重于将经典渲染原理与学习的3D场景表示（通常现在称为**神经场景表示**）相结合的方法。这些方法的一个关键优点是，设计上的3D一致，从而实现了捕获场景的新视点合成等应用。除了处理静态场景的方法外，还介绍了用于建模**非刚体变形目标** 的神经场景表示以及场景编辑和合成。虽然这些方法大多是场景特定的，但也讨论了跨目标类进行泛化的技术，并可用于生成任务。除了回顾这些最先进的方法，还概述了使用的基本概念和定义。最后，讨论了公开挑战和社会影响。

虽然传统的计算机图形学允许生成场景的高质量可控图像，但场景的所有物理参数（例如，摄像机参数、照明和对象的材质）都需要作为输入提供。如果想要生成真实场景的可控图像，需要从现有的观测（如图像和视频）中估计这些物理属性，即逆渲染，非常具有挑战性，特别是当目标是照片级真实的合成图像。

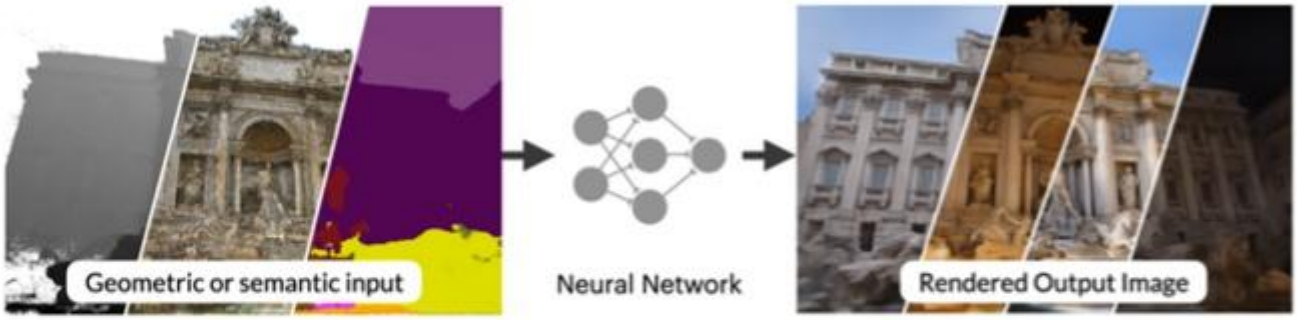
相比之下，神经渲染是一个迅速兴起的领域，它允许场景的紧凑表示，并且可以通过神经网络从现有观测中学习渲染。神经渲染的主要思想是结合经典（基于物理的）计算机图形学的见解和深度学习的最新进展。与经典计算机图形学类似，神经渲染的目标是以可控的方式生成照片级真实感图像，例如，新视点合成、重照明、场景变形和合成等。

这方面的一个很好的例子是最近的神经渲染技术，该技术试图通过仅学习3D场景表示并依赖计算机图形中的渲染函数进行监督来分离建模和渲染过程。例如，**神经辐射场（NeRF）** 使用多层感知器（MLP）来近似3D场景的辐射场和密度场。该学习的体表示可以使用解析可微分渲染（即体积分）从任何虚拟摄像头渲染。对于训练，假设从多个摄像机视点观测场景。从这些训练视点，渲染估计的3D场景，并最小化渲染图像和观察图像之间的差异，根据这些观察结果训练网络。一旦训练完成，由神经网络近似的3D场景可以从新的视点进行渲染，从而实现可控合成。与使用神经网络学习渲染函数的方法相反，NeRF在该方法中更明确地使用了计算机图形学的知识，由于（物理）归纳偏差，能够更好地概括新视图：场景密度和半径的中间3D

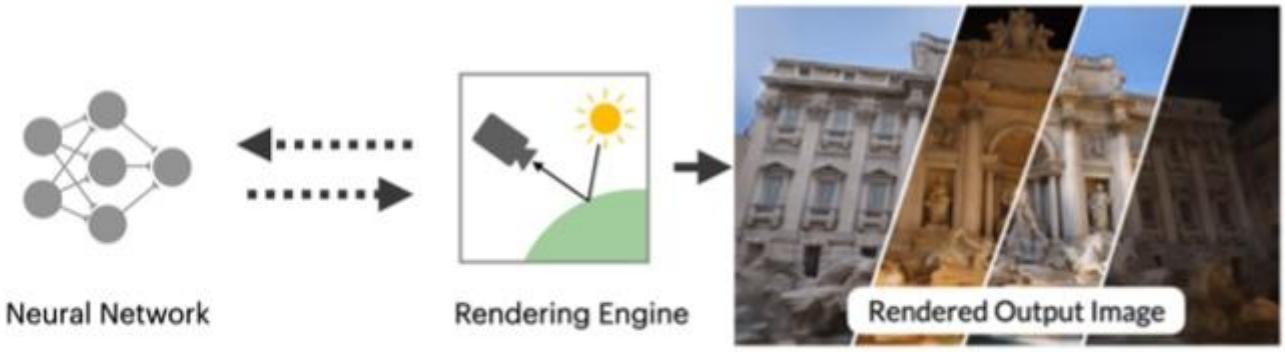
结构化表示。因此，NeRF在3D空间中学习物理上有意义的颜色和密度值，物理激发的光线投射和体集成可以持续渲染到新视图中。

所取得的结果质量，以及方法的简单性，导致了该领域的“爆炸式”发展。已经取得了一些进步，这些进步提高了适用性，实现了可控性，动态变化场景的捕获以及训练和推理时间。由于神经渲染是一个发展非常快的领域，在许多不同的维度上都取得了重大进展，因此对最近的方法及其应用领域进行了分类，以提供发展的简要概述。

在本报告中，重点介绍了将经典渲染与可学习3D表示相结合的高级神经渲染方法（见图）。



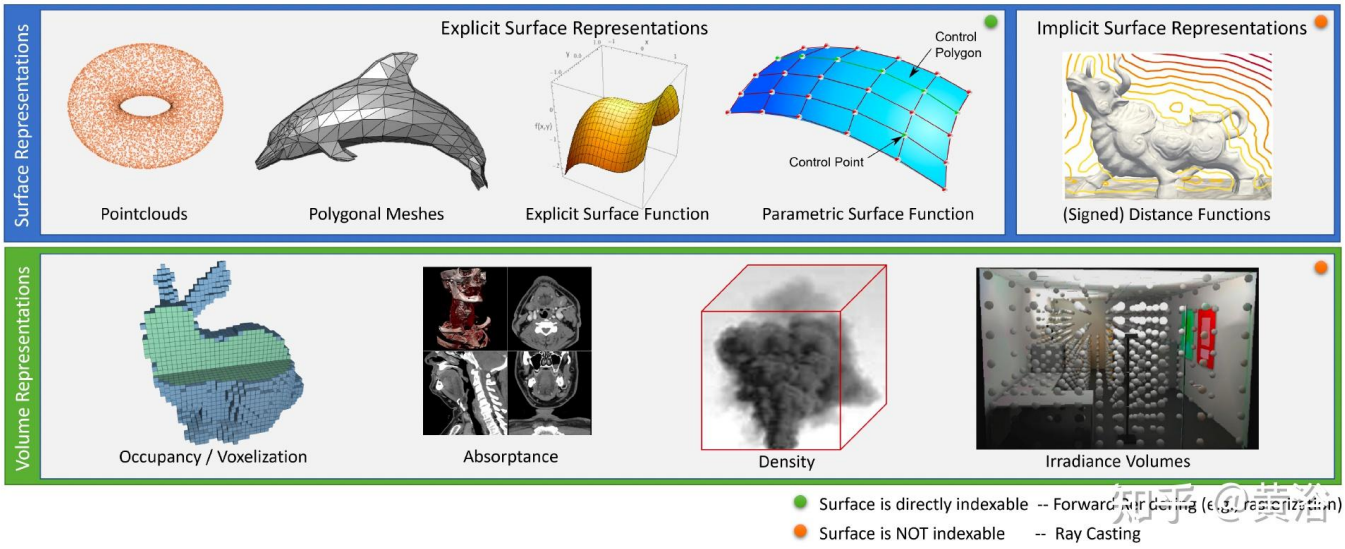
(a) 2D Neural Rendering, also known as neural refinement, neural re-rendering, or deferred neural rendering is based on 2D inputs that are generated for example using a classical renderer and *learns to render a scene in 2D*.



(b) 3D Neural Rendering *learns to represent a scene in 3D* and uses fixed differentiable rendering schemes from computer graphics which are motivated by physics.

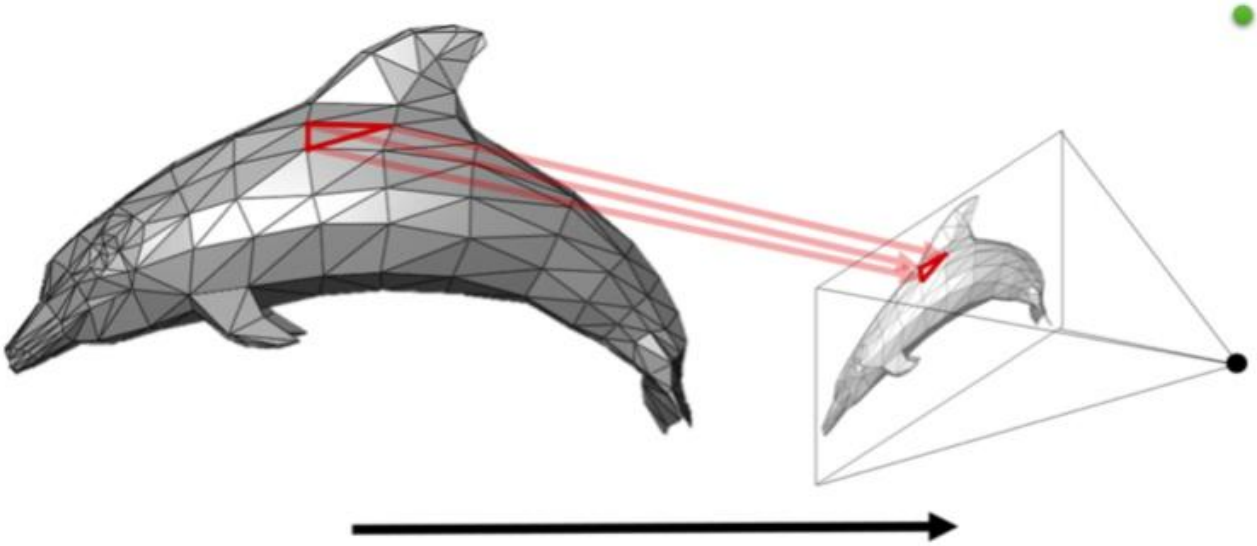
基础的神经3D表示在设计上是3D一致的，并能够控制不同的场景参数。在本报告中，全面概述了不同的场景表示，并详细介绍了从经典渲染流水线以及机器学习中借鉴的组件基本原理。进一步关注用神经辐射场以及体渲染的方法。然而，这里忽略主要在2D屏幕空间中推理的神经渲染方法，也不包括光线跟踪图像的神经超采样和去噪方法。

几十年来，计算机图形学界探索了各种表征，包括点云、隐式和参数曲面、网格和体积（见图）。

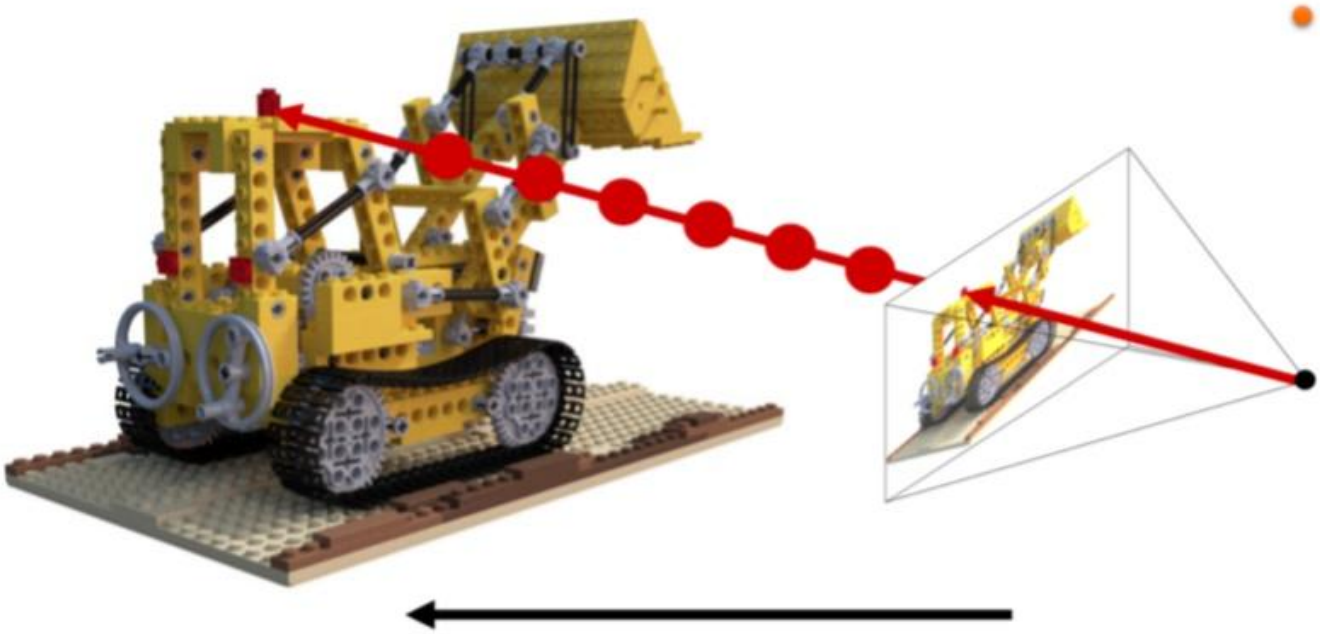


虽然这些表示在计算机图形学领域有明确的定义，但当前神经渲染的文献中经常存在混淆，特别是当涉及到隐式和显式曲面表示和体积表示。通常，体表示可以表示曲面，但反之亦然。体表示存储体特性，如密度、不透明度或占用率，但它们也可以存储多维特征，如颜色或亮度。与体表示不同，曲面表示存储目标曲面的特性。它们不能用于模拟体物质，如烟雾（除非是粗略近似值）。对于曲面和体表示，都有连续和离散的对应项（见上图）。连续表示对于神经渲染方法特别有趣，因为它们可以提供解析梯度。

将3D场景渲染为2D图像平面有两种常用方法：**光线投射和光栅化**，参见下图。还可以通过在场景中定义摄像头来计算场景的渲染图像。大多数方法使用针孔摄像头，其中所有摄像头光线通过空间的单个点（焦点）。对于给定的摄影机，可以将来自摄影机原点的光线投射到场景，计算渲染图像。



(a) Forward Rendering (e.g., rasterization) – the image is generated by projecting the 3D representation to the image plane.



(b) Ray Casting – the image is generated by casting viewing rays, sampling the 3D representation and accumulating them. Image adapted

要正确建模当前摄像机图像需要考虑镜头。撇开景深或运动模糊等必须在图像形成过程中建模的影响不谈，还给投影函数增加失真效应。不幸的是，没有一个简单的模型来捕捉所有不同的镜头效果。标定包，如OpenCV给的，通常实现具有多达12个失真参数的模型。它们通过五次多项式建模，因此不是简单

可逆的（这是光线投射所需要的，而不是点投影）。更现代的摄像机标定方法使用了更多的参数，实现了更高的精度，并且可逆和可微分。

直接光栅化主要用网格，网格由一组顶点v和面f描述，连接三个或四个顶点以定义曲面。一个基本的观点是，3D中的几何操作只能处理顶点：例如，用相同的外参矩阵E将世界中的每个点变换到摄像头坐标系。转换后，可以剔除视锥体以外的点或具有错误法线方向的点，减少下一步要处理的点-面的数量。投影到图像坐标的点位置也可以通过内参矩阵K轻松找到。表面信息可用于插值表面基元的深度，最上面的表面可存储在z-缓冲区中。但是，很难以此捕获某些效果（例如，照明效果、阴影、反射）。它可以通过“软”光栅化技术进行细分。

下面按照应用来讨论神经渲染和神经场景表征的各种方法：静态场景的新视点合成、对目标和场景的泛化、非静态场景的视角合成、场景编辑和组合、重照明和材料编辑等。

1 新视图合成

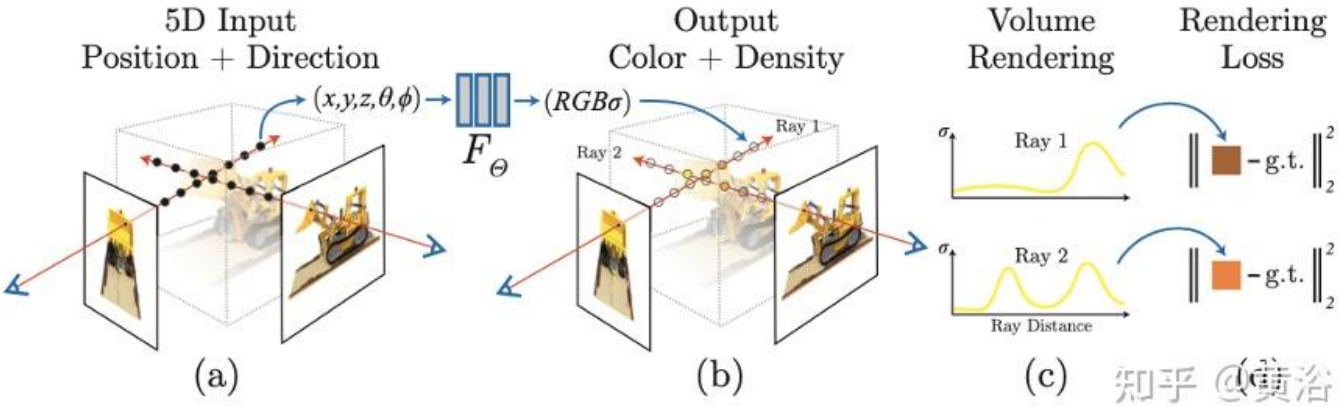
新视图合成是在给定一组图像及其摄像头姿态作为输入的情况下，从新的摄像头位置渲染给定场景。

视图合成方法基于几个重要标准进行评估。显然，输出图像应尽可能逼真。然而，这并不是全部，也许更重要的是多视图3D一致性。当摄像机在场景中移动时，渲染视频序列必须显示为描绘一致的3D内容，而不会闪烁（flickering） 或扭曲

（warping）。随着神经渲染领域的成熟，大多数方法都朝着生成固定3D表示的方向发展，该输出可用于渲染新的2D视图。这种方法自动提供了一定程度的多视图一致性，当过度依赖黑盒的2D卷积网络作为图像生成器或渲染器时，过去一直很难实现这种一致性。

为了解决体素网格分辨率和内存限制，Scene Representation Networks（SRN）将基于球体跟踪的神经渲染器与多层感知器（MLP）相结合，作为场景表示，主要关注场景的泛化，实现少镜头的重建。Differentiable Volumetric Rendering（DVR）类似地利用了表面渲染（surface rendering）方法，但证明了单场景的过拟合能够重建更复杂的外观和几何体。

神经辐射场（NeRF） 标志着将基于MLP的场景表示应用于单场景、照片级真实感新视图合成的突破，见下图。



与基于表面的方法不同，NeRF直接应用体渲染（volume rendering）模型，从MLP合成图像，从输入位置和观察方向映射到输出体密度和颜色。基于输入图像的像素级渲染损失，优化一组不同的MLP权重来表示每个新输入场景。

基于MLP的场景表示由于在优化过程中有效地对场景进行了可微分压缩，因此实现比离散3D体更高的分辨率。例如，渲染800×800分辨率输出图像的NeRF表示能力仅需要5MB的网络权重。相比之下，800^3 的RGBA体素网格将消耗接近2GB的存储空间。

这种能力可归因于，NeRF在通过MLP之前对输入空间坐标应用位置编码。与之前使用神经网络表示隐式曲面或者隐式体的工作相比，NeRF的MLP能表示高得多的频率信号，而不增加其容量（根据网络权重的数量）。

从离散3D网格切换到基于MLP的表示的主要缺点是渲染速度。要计算空间中单个点的颜色和密度，而不是直接查询简单的数据结构，需要评估整个神经网络（数十万次浮点操作）。在典型的台式机GPU上，标准深度学习框架中实现NeRF渲染单个高分辨率图像需要几十秒。

有一些加速基于MLP表征的体渲染方法提出，比如Neural Sparse Voxel Fields和KiloNeRF。还有几个方法在稀疏3D网格上缓存NeRF MLP学习的各种量，允许在训练完成后进行实时渲染，比如SNeRG，FastNeRF，PlenOctrees和NeX-MPI等。加速渲染的另一种方法是训练MLP表示本身，有效地预计算沿光线的部分或全部体积分，如AutoInt和Light Field Networks。

很多新方法采用网格、稀疏网格、树和哈希等经典数据结构，加速渲染速度，实现更快的训练时间。Instant Neural Graphics Primitives利用多分辨率哈希编码，而不是显式网格结构，在几秒钟内实现NeRF的训练。

还有一些改进包括监督数据（比如深度值）、最优化摄像头姿态、混合表面/体表征、鲁棒性和质量改进（NeRF++、MipNeRF）、NeRF和标准计算成像方法的结合（Deblurr-NeRF、NeRF in the Dark、HDR-NeRF和NeRF-SR等）、大规模场景和来自文字的NeRF（Dream NeRF和CLIP NeRF）等。

---

2 目标和场景的泛化

大量的工作涉及基于体素、基于网格或非3D结构化神经场景表示的多场景和目标类的泛化，这里主要讨论基于MLP的场景表示做泛化的最新进展。其中，在单个场景中过拟合单个MLP的方法需要大量的图像观测数据，在场景表示中推广的核心目标是在给定很少或可能只有单个输入视图的情况下进行新视图合成。概述中方法分类如下：是否利用局部或全局条件，是否可以用作非传统生成模型，利用什么样的3D表示（体积、SDF或占用），需要什么样的训练数据，以及如何执行推理（通过编码器、自动解码器框架或基于梯度的元学习等）。

有两种关键的方法来概括不同的场景。一类工作遵循了一种类似于基于图像渲染（IBR）的方法，其中多个输入视图被扭曲（warp）和混合（blend）合成新的视点。在基于MLP的场景表示上下文中，这通常通过局部调节实现，其中场景表示MLP的坐标输入与存储在离散场景表示（比如体素网格voxel grid）中的局部变化特征向量连接一起。

PiFU使用图像编码器计算输入图像的特征，并通过在图像平面投影3D坐标来调整这些特征的3D MLP。然而，PiFU没有可微分渲染器，因此需要真值3D监督。PixelNeRF和Pixel-Aligned Avatars 在体渲染框架中利用此方法，其中这些特征在多个视图上聚合，MLP生成颜色和密度场，以NeRF方式渲染。当在多个场景上进行训练时，可学习用于重建的场景先验，从几个视图中高保真地重建场景。

PixelNeRF还可以在特定的目标类上进行训练，从而能够从一个或多个姿态图像进行目标实例3D重建。GRF使用了类似的框架，带有一个额外的注意模块，用于说明在不同采样输入图像中3D点的可见性。Stereo Radiance Fields类似地从多个上下文视图中提取特征，但利用上下文图像成对特征之间的学习对应匹配来聚合上下文图像间的特征，而不是简单的平均聚合。最后，IBRNet和NeR-Former在光线采样中引入transformer网络，推理可见性。LOLNeRF学习仅用单目监督的肖像图像广义NeRF模型。生成器网络以实例特定的潜向量为条件，进行联合训练。GeoNeRF构建了一组级联成本体，并使用transformer推断几何结构和外观。

一种基于图像方法的替代方案旨在学习场景的整体、全局表示，而不是依赖图像或其他离散空间数据结构。给定一组观测，其实现通过为场景表示MLP推断一组权重来描述整个场景。一些工作在单个低维潜代码中编码场景来实现，然后用该代码调节场景表示MLP。

Scene Representation Networks（SRN）通过超网络将低维潜代码映射到MLP场景表示的参数，然后通过光线行进（ray-marching）渲染生成的3D MLP。为了重建给定姿态视图的实例，SRN优化潜代码，其渲染与输入视图匹配。Differentiable Volumetric Rendering类似使用表面渲染，解析计算其梯度，并通过CNN编码器执行推断。Light Field Networks利用低维潜代码直接参数化3D场景的4D光场，实现单次评估渲染。

NeRF VAE将NeRF嵌入到variational auto-encoder （VAE），类似地在单个潜代码中表示整个场景，但学习一个生成模型以实现采样. Sharf采用在一个类中目标体素化形状的生成模型，这继而调节了一个更高分辨率的神经辐射场，其采用体渲染，获得更高的新视图合成保真度。

潜代码为条件，Fig-NeRF将目标类别建模为模板形状，该潜代码经历相同潜变量为条件的变形。这使得网络能够将某些形状变化解释为更直观的变形。Fig-NeRF着重于从真实目标扫描中检索目标类别，也提出用学习的背景模型从其背景分割目标。一种替代是将场景表示为低维潜代码，通过基于梯度的元学习，在几个优化步骤中快速优化MLP场景表示的权重. 这可用

于从少量图像快速重建神经辐射场. 当在新场景训练时，预训练模型收敛速度更快，与标准神经辐射场训练相比，需要更少的视图。

Portrait-NeRF 提出了一种元学习方法来从人的单个正面图像中恢复NeRF。为了说明受试者之间姿态的差异，在姿势不可知（pose-agnostic）的标准参考框架中建模3D肖像，用3D关键点对每个受试者进行扭曲。利用基于梯度的元学习和图像特征上的局部调整可快速恢复场景的NeRF。

可以利用类似方法学习无条件生成模型，而不是根据寻找3D场景的一组观测来推断低维潜代码。这里，配备有神经渲染器的3D场景表示嵌入到生成对抗网络（GAN）中。不是从一组观测值推断低维潜代码，而是定义潜代码的分布。在前向过程中，从该分布中采样一个潜变量，调整MLP场景表示，并通过神经渲染器渲染图像。该图像可用于对抗性损失。仅给定2D图像，这能够学习3D场景形状和外观的3D生成模型。通过体素网格参数化3D场景表示的框架，**GRAF** 首先在其中利用条件NeRF，并在photorealism中实现显著的改进。Pi-GAN通过一个SIREN（“Implicit neural representations with periodic activation functions“）结构的基于FiLM（“Film: Visual reasoning with a general conditioning layer“.）调节方案进一步改进了该架构。

最近一些方法探索了提高这些生成模型质量和效率的不同方向。几何重建的计算成本和质量可以通过表面表示法来提高。除了为鉴别器合成多视图图像外，ShadeGAN使用显式着色步骤也在不同照明条件下生成输出图像渲染，用于更高质量的几何重建。混合技术方面探索了许多方法，其中基于图像的CNN网络用于优化3D生成器的输出。图像空间网络能够以更高的分辨率和更高保真输出进行训练。一些方法探索将生成模型分解为单独的几何和纹理空间。这里，一些方法学习图像空间中的纹理，而其他方法在3D中同时学习几何和纹理。

虽然这些方法不需要每个3D场景进行一次以上的观测，也不需要摄像机姿态真值，但仍然需要摄像机姿态分布的知识（对于肖像图像，摄像机姿态分布必须产生合理的肖像角度）。CAMPARI通过联合学习摄像头姿态分布和生成模型来解决这一约束。GIRAFFE提出将场景参数化为多个前景（目标）NeRF和单个背景NeRF的组合来学习由多个目标组成的场景生成模型。对每个NeRF单独采样潜代码，并由体渲染器将其合成为合理2D图像。

3 动态场景的扩展

原始神经辐射场用于表示静态场景和目标，还有一些方法可以额外处理动态变化的内容。这些方法可以被分类为**时变表征** 方法，允许将动态变化场景的新视点合成为未经修改的回放（例如，产生子弹-时间效果），或者也可以分类为**控制变形状态** 的技术，其允许对内容进行新视点合成和编辑。变形的神经辐

射场可以隐式或显式实现，见图所示：左边是隐式地实现，在变形（时间t）上调节辐射场v。右边为显式地实现，用单独的变形MLP去扭曲（warp）空间，回归从变形空间（黑色）到静态规范空间（黄色）的偏移（蓝色箭头）。这个变形将直线光线弯曲到标准辐射场。

- **时变表征**

时变NeRF允许播放具有新视点的视频。由于放弃了控制，这些方法不依赖于特定的运动模型，因此可以处理一般目标和场景。

同时一些工作提出了非刚性场景NeRF的若干扩展。首先讨论隐式模拟变形的方法。虽然原始NeRF是静态的，仅将3D空间点作为输入，但可以简单的方式扩展为时变的：另外体表征可以取决于表示变形状态的向量。在当前的方法中，这种调节采用时间输入（可能位置编码）或每个时间步长的自动解码潜代码。

没有目标类型或3D形状先验知识的情况下，处理非刚性场景是一个不适定问题，这类方法采用了各种几何正则化方法，以及在附加数据模式上的条件学习。为了鼓励反射和不透明度在时间上的一致性，有几种方法学习时域相邻时间步长之间的**场景流映射** . 由于这仅限于小的时间邻域，因此无畸变的新视图合成主要在接近时空输入摄像机轨迹演示。

场景流映射可以用重建损失进行训练，重建损失将场景从其他时间步长扭曲到当前时间步长，其鼓励估计光流与场景流2D投影之间的一致性，或3D跟踪逆投影的关键点。场景流通常受到额外正则化损失的约束，例如鼓励空间或时间平滑性或前向-后向循环一致性。与提到的其他方法不同，**Neural Radiance FLow（NeRFlow）** 对具有无穷小位移的变形进行建模，其需要和Neural ODE集成获得偏移估计。

此外，一些方法使用估计的深度图来监督几何估计。这种正则化的一个限制是重建的精度取决于单目深度估计方法的精度。因此，在新视图中可以看到单目深度估计方法的伪影。

最后，静态背景通常是单独处理的，允许时域单目输入的多视图线索。为此，一些方法估计不以变形为条件的第二静态体，或者引入软正则化损失来约束静态场景内容。

NeRFlow 可用于预训练场景的去噪和超分辨率视图。NeRFlow的局限性包括：难以保持静态背景、处理复杂场景（非分段刚性变形和运动）以及在与输入轨迹基本不同的摄像头轨迹下渲染新视图。

迄今为止，出现的方法用取决于变形的场景表示隐式地建模变形。这使得变形的控制变得繁琐和困难。其他工作将变形与几何和外观分离：将变形分解为静态规范场景之上的独立函数，这是实现可控性的关键一步。变形的实现是将直线射线投射到变形空间，并弯曲到规范场景中，通常是用基于坐标的MLP回归直线射线上的点偏移。这可以被认为是空间扭曲或场景流。

与隐式建模相反，这些方法通过静态标准场景的构造，在时间上共享几何和外观信息，从而提供不会漂移的硬对应关系。由于该硬约束，与隐式方法不同，具有显式变形的方法无法处理拓扑变化，并且仅在运动明显小于采用隐式方法的场景，才能演示结果。

D-NeRF 使用无正则化的光线弯曲（ray-bending）MLP来模拟从背景分割出的单个或多个合成目标变形，通过虚拟摄像机观察。其假设给定一组预定义的多视图图像，但在训练时只有选择一个单视图用于监督。因此，D-NeRF可被视为多视角监督技术和真正单目监督方法之间的中间步骤。

一些工作展示了移动单目摄像头观察的真实场景结果。Deformable NeRF的核心应用是Nerfies的构建，即自由视点自拍。Deformable NeRF用每个输入视图的自动解码潜代码来调节变形和外观。弯曲光线用尽可能刚性的项（也称为弹性能量项）进行正则化，惩罚与分段刚性场景配置的偏差。

因此，Deformable NeRF在铰接场景（例如，一只手拿网球拍）和包括人头部的场景（其中头部相对躯干移动）中工作良好。尽管如此，小的非刚性变形处理得很好（如微笑），因为正则化子是软的。这项工作的另一个重要创新是从粗到细（from-coarse-to-fine）方案，允许首先学习低频分量，并避免由于过拟合高频细节而导致的局部极小值。

HyperNeRF 是Deformable NeRF的延伸，用一个规范超空间而不是一个单规范框架。这允许处理具有拓扑变化的场景，例如张开和闭合嘴巴。在HyperNeRF中，Deformable NeRF的bending network（MLP）通过一个周围切片表面网络（同样是MLP）来增强，其间接调节变形的规范场景，为每个输入RGB视图选择一个规范子空间。因此，它是一种结合显式和隐式变形建模的混合模型，允许牺牲硬对应关系来处理拓扑变化。

Non-rigid NeRF （NR NeRF）用场景规范体、场景刚性标志（MLP）和帧光线弯曲算子（MLP），对时变场景外观进行建模。NR NeRF表明，处理具有较小非刚性变形和运动的场景不需要额外的监督提示，如深度图或场景流。此外，观察到的变形由一个发散算子来正则化，施加体保持（volume-preserving ）约束，相对于监督单目输入视图可以稳定遮挡区域。在这方面，它与Nerfies的弹性正则化子具有类似的性质，后者惩罚分段刚性变形的偏差。这种正则化使得新视图的摄像机轨迹与输入摄像机轨迹显著不同。虽然可控性仍然受到严重限制，但NR NeRF演示了对学习的变形场的几个简单编辑，例如运动放大或动态场景内容移除。

其他方法并不局限于单目RGB输入视频的情况，而是考虑存在其他输入。

Time-of-Flight Radiance Field s（TöRF）方法替换数据驱动的先验知识，用来自深度传感器的深度图重建动态内容。与绝大多数计算机视觉工作不同，TöRF使用原始ToF传感器测量（所谓的phasors），这在处理弱反射区域和现代深度传感器的其他限制（例如，受限的工作深度范围）时带来了优势。在NeRF学习中，集成测量的场景深度减少了对输入视图数量的要求，从而产生清晰和详细的模型。与NSFF和时空神经辐射场相比，深度提示可提供更高的精度。

Neural 3D Video Synthesis 用多视图RGB设置并隐式地建模变形。该方法首先对关键帧进行训练，利用时间平滑性。它还设定摄像机保持静态，场景内容大部分静态，通过有偏差的方式采样光线进行训练。即使对于较小的动态内容，结果也很清晰。

- **控制变形状态**

为了控制神经辐射场的变形，这类方法使用特定类别的运动模型作为变形状态的基本表示（例如，人脸的变形模型或人体的骨骼变形图）。

NeRFace 是第一种使用可变形模型隐式控制神经辐射场的方法。他们使用面部跟踪器在训练视图（单目视频）中重建人脸混合形状参数和摄像机姿态。MLP在这些视图上用混合形状参数和可学习的每帧潜代码作为条件进行训练。此外，它们假设已知的静态背景，使辐射场仅存储关于面部的信息。潜代码用于补偿丢失的跟踪信息（即人的肩膀）以及跟踪中的错误。训

练后，可以通过混合形状参数控制辐射场，从而允许复现（reenactment）和表情编辑。

一种受NeRFace启发的音频驱动神经辐射场（**AD-NeRF** ），不用表情系数，而是Deep-Speech提取的音频特征映射到给辐射场表征MLP提供条件的一个特征。虽然表情是通过音频信号隐式控制的，但提供了对头部刚性姿态的显式控制。为了合成一个人的肖像视图，他们使用两个单独的辐射场，一个用于头部，一个为躯干。

“I M Avatar”基于皮肤场扩展NeRFace，其用于在给定新的表情和姿态参数的情况下使规范NeRF体变形。

除了这些特定主题的训练方法外，**Head-NeRF** 和**MoFaNeRF** 提出了一种广义模型，用于表示不同视图、表情和照明下的人脸。与NeRFace类似，它们根据控制如人物形状、表情、反照率和照明等附加参数来调节NeRF MLP。这两种方法都需要细化网络（2D网络）来改进基于条件NeRF MLP做体渲染的粗略结果。

虽然上述方法在肖像场景中显示了有希望的结果，但它们不适用于高度非刚性变形，尤其是从单个视图捕获的人体铰接（articulated）运动。因此，需要显式地利用人体骨架嵌入。**Neural Articulated Radiance Field（NARF）** 通过姿态标注图像进行训练。铰接目标分解为多个刚性目标部分，其局部坐标系和全局形状变化位于之上。收敛的NARF可通过操纵姿态、估计深度图和执行身体部位分割来渲染新视图。

与NARF相比，**A-NeRF** 以自监督的方式从单目镜头中学习演员特定的体神经人体模型。该方法将动态NeRF体与关节式人体骨骼嵌入的显式可控性相结合，并以合成解析的方式重建姿态场和辐射场。一旦训练过，辐射场可以用于新视点合成以及运动重定位。

当A-NeRF在单目视频上训练时，**Animatable Neural Radiance Fields（ANRF）** 是一种从多视图视频重建人体模型的骨架驱动（skeleton- driven）方法。其核心组件是一种新运动表示，即神经混合权重场，其与3D人体骨骼相结合，用于生成变形场。类似于几种通用非刚性NERF，ANRF保持一个规范空间，并估计多视图输入和规范框架之间的双向对应关系。

重建的可动画人体模型可用于任意视点渲染和新姿态下的重渲染。通过在离散化正则空间点的体密度上运行移动立方体（marching cubes）算法，也可以从ANRF中提取人体网格。该方法实现了所学习人体模型的高视觉精度，在未来的工作中，可以改进处理观察表面的复杂非刚性变形（如因宽松衣服引起的变形）。

Neural Body 方法能够从稀疏多视图视频（例如，四个同步视图）中实现人类表演的新视图合成。他们的方法通过参数化人体形状模型SMPL进行调节，作为一个形状智体先验。它假设不同帧所恢复的神经表示具有锚定到一个可变形网格的相同潜代码集。通用基线，如rigid NeRF （每个时间戳应用）或Neural Volumes假设更致密的输入图像集。因此，从几个同步

输入图像呈现运动人体的新视图，都无法与Neural Body竞争。该方法还与人类网格重建技术（如PIFuHD进行了有利的比较，当涉及精细外观细节（例如，很少穿或独特的服装）的3D重建时，该技术强烈依赖于训练3D数据。

类似于Neural Body方法，**Neural Actor（NA）** 和**HVTR** 用SMPL模型来表示变形状态。他们利用智体显式地将周围的3D空间解缠绕成标准姿势，其中嵌入了NeRF。为了改善几何和外观高保真细节的恢复，他们使用在SMPL曲面上定义的附加2D纹理图，作为NeRF MLP的附加条件。

H-NeRF 是另一种利用人体模型条件做时域三维重建的技术。类似于Neural Body，它们需要来自同步和标定摄像机的稀疏视频集。与之相反，H-NeRF使用符号距离场的结构化隐式人体模型，带来更清晰的渲染和更完整的几何结构。与H-NeRF类似，DD-NeRF构建在符号距离场之上，渲染整个人体。给定多视图输入图像和重构的SMPL体，它们用体渲染累积的回归SDF和辐射值。

Human-NeRF 也基于输入的多视图，但学习用于任意视点渲染的广义神经辐射场，其可针对特定演员进行微调。另一项工作叫做HumanNeRF，用一个通用非刚性运动场细化的骨架驱动运动场，展示了如何基于单目输入数据训练特定演员的神经辐射场。

Mixture of Volumetric Primitives 用于实时渲染动态、可动画的虚拟人模型。主要思想是用一组可以动态改变位置和内容的体基元，对场景或目标建模。这些基本体基元，像基于零件的模型一样，为场景的组件建模。每个体基元是由解码器网络从潜代码生成的体素网格。该潜代码定义了场景的配置（例如，人脸的情况下的面部表情），解码器网络用该配置来生成原始位置和体素值（包含RGB颜色和不透明度）。

要渲染，用一个光线行进（ray marching ）程序沿着每个像素对应的光线累积颜色和不透明值。与其他动态NeRF方法类似，多视图视频被用作训练数据。该方法能够创建非常高质量的实时渲染，即使在具有挑战性的材质（如头发和衣服）上也看起来逼真。**E-NeRF** 则展示了基于深度引导采样技术的高效NeRF渲染方案。它们使用多视图图像作为输入，演示移动人体和静态目标的实时渲染。

4 组合和编辑

迄今为止讨论的方法允许重建静态或动态场景的体表示，并可能从几个输入图像渲染它们的新视图。保持观测的场景不变，除了相对简单的修改（例如前景移除）。最近的几种方法也允许编辑重建的3D场景，即重新安排和仿射变换目标，并改变其结构和外观。

Conditional NeRF 可以通过手动用户编辑改变2D图像中所观测刚性目标的颜色和形状（例如，可以移除一些目标部分）。该功能在同一类多个目标实例训练的单NeRF启动。在编辑过程中，调整网络参数匹配新观察实例的形状和颜色。这项工作的贡献之一是找到了可调参的子集，可以成功传播用户编辑生成新视图。这避免了对整个网络进行昂贵的修改。**CodeNeRF** 表示目标类中的形状和纹理变化。与**pixelNeRF** 类似，CodeNeRF可以合成未见过目标的新视图。它学习形状和纹理的两种不同嵌入。在测试时，它从单个图像中估计摄像机姿态、目标的3D形状和纹理，并且可以通过改变潜代码来连续修改。Co-deNeRF在不假设已知摄像机姿态的情况下，实现了与先前单图像3D重建方法相当的性能。

Neural Scene Graphs（NSG） 是一种从驾驶录取的单目视频（自车视图）合成新视图的方法。该技术将多个独立刚性移动目标的动态场景分解为学习的场景图，该场景图对单个目标变换和辐射进行编码。因此，每个目标和背景由不同的神经网络编码。此外，静态节点的采样被限制为分层面（其平行于图像平面）提高效率，即2.5D表示。NSG要求输入帧集合上每个刚性运动的感兴趣目标的标注跟踪数据，并且每个目标类别（例如，汽车或公共汽车）共享单个体先验。然后，可以使用神经场景图来渲染相同（即，观察到的）或编辑（即，通过重新排列对象）场景的新视图。NSG的应用包括背景-前景分解、丰富汽车感知的训练数据集以及改进目标检测和场景理解。

另一种分层表示，在空间和时间上一致的NeRF（**ST-NeRF** ）依赖于所有独立移动和铰接目标的边框，从而产生多个层，并解开它们的位置、变形和外观信息。ST-NeRF的输入是一组16个的同步视频，这些视频来自以规则间隔放置在半圆中的摄像机，以及人类背景分割掩码。该方法的名称表明，时空一致性约束反映在其架构中，即作为规范空间的时空变形模块和NeRF模块。ST-NeRF也接受时间戳来解释外观随时间的演变。渲染新视图时，采样光线会投射到多个场景层，这会导致累积密度和颜色。ST-NeRF可用于神经场景编辑，如重缩放、移动、复制或移除表演者，以及时间重安排。

5 重照明和材料编辑

以上应用基于简化的吸收-发射体渲染模型，其中场景被建模为阻挡和发射光的粒子体。虽然该模型足以从新视点渲染场景的图像，但它不能在不同的照明条件下渲染场景的图片。启用重照明需要场景表示，该场景表示可以模拟光通过体的传输，包括具有各种材质属性的粒子对光的散射。

Neural Reflectance Fields 建议首次扩展NeRF以实现重照明。与NeRF不同，Neural Reflectance Fields不是将场景表示为体密度场和与视图相关的辐射亮度场，而是将场景表示成体密度场、表面法线和双向反射分布函数（BRDF）。这允许在任意照明条件下渲染场景，方法是使用每个3D位置处的预测曲面法线和BRDF来评估该位置处的粒子向摄像头反射了多少入射光。然而，对于神经体渲染模型，评估从沿摄像头光线的每个

点到每个光源的可见性在计算上非常紧张。即使仅考虑直接照明，MLP也必须在沿摄像头光线和每个光源的每个点之间的密集采样位置进行评估，以便计算入射照明，对该光线进行渲染。神经反射场仅用摄像头同定位的单点光照明，这样得到的目标图像进行训练可回避此问题，因此只需要沿摄像头光线评估MLP。

恢复可重照明模型的其他近期工作，简单地忽略自遮挡，并假设任何表面上方上半球的所有光源都是完全可见的，避免了计算光源可见性的困难。两个方法**PhySG** 和**NeRD** ，假设全光源可见性，并将环境照明和场景BRDF表示为球面高斯的混合来进一步加速渲染，使入射光乘以BRDF的半球积分，能够以封闭形式计算。假设全光源能见度可以很好地适用于大多数凸形的目标，但该策略无法模拟场景几何体遮挡光源而产生的效果，如投射阴影。

Neural Reflectance and Visibility Fields (NeRV) 训练MLP以近似任何输入3D位置和2D入射光方向的光源可见性。与沿每条光线的密集采样点处查询MLP不同，这里对于每个入射光方向，只需查询一次可见性MLP。这使得神经网络能够从具有显著阴影和自遮挡效果的图像中恢复场景的可重照明模型。

与之前讨论的方法不同，**NeRFactor** 从预训练的NeRF模型开始。然后，NeRFactor将预训练NeRF的体几何简化为曲面模型，优化MLP表示曲面上任意点的光源可见性和曲面法线，并最终优化任意曲面点的环境照明和BRDF表示，恢复可重照明模

型。这导致在渲染图像时更有效的可重照明模型，因为体积几何已简化为单个曲面，并且可以通过单个MLP查询计算任意点的光源可视性。

NeROIC 技术还使用多级流水线，从多个无约束照明环境下捕获的目标图像中恢复可重照明的NeRF类模型。第一阶段恢复几何，同时解释由于具有潜外观嵌入的照明引起的外观变化，第二阶段从恢复的几何中提取法向量，第三阶段估计BRDF特性和照明的球面谐波表示。

与上述侧重于恢复目标的可重照明表示不同，**NeRF-OSR** 恢复大型建筑和历史遗址的NeRF式可重照明模型。NeRF OSR采用Lambertian模型，并将场景分解为漫射反照率、表面法线、照明的球面谐波表示和阴影，这些结合起来在新的环境照明下场景重照明。

上述可重照明模型将场景材质表示为BRDF的连续3D场。这将启用一些基本的材质编辑，因为可以在渲染之前更改恢复的BRDF。**NeuTex** 引入曲面参数化网络，学习从体中3D坐标到2D纹理坐标的映射，从而实现更直观的材质编辑。恢复场景的NeuTex模型后，可以轻松编辑或替换2D纹理。

Ref-NeRF 专注于提高NeRF表示和渲染镜面反射曲面的能力。虽然Ref-NeRF不能用于重照明，因为它不能将入射光与反射特性分离开来，但它将发射光构造为物理意义上的组件（漫反射和镜面反射颜色、法线向量和粗糙度），从而支持直观的材质编辑。

6 光场

体渲染、球跟踪和其他3D渲染前向模型可以产生照片逼真的结果。然而，对于给定光线，它们都需要在光线第一次与场景几何相交的任何3D坐标处对底层3D场景进行采样。由于该交点先验未知，光线行进（ray-marching）算法首先必须发现该曲面点。最终，这会产生与场景几何复杂性成比例的时间和内存复杂性，其中必须对越来越多的点进行采样以渲染越来越复杂的场景。在实践中，每射线有数百甚至数千个点。此外，精确渲染反射和二阶照明效果需要多次反弹光线跟踪，因此对于每个像素，必须跟踪多条光线，而不是仅跟踪一条光线。这产生了高计算负担。虽然在重建单个场景（过拟合）的情况下，这可以通过机灵数据结构、哈希和专家低层工程来避免，但在重建仅给出少量观察或甚至仅给出单个图像的3D场景情况下，这种数据结构阻碍了学习重建算法的应用，例如使用卷积神经网络从单个图像推断3D场景的参数。

7 工程框架

使用神经渲染模型给从业者带来了显著的工程挑战：大量图像和视频数据必须以高度非顺序的方式处理，模型通常需要区分大型和复杂的计算图。开发高效的算子通常需要使用低级语言，这同时使得使用自动区分更加困难。一些工具的最新进展，可以帮助克服与神经渲染相关的整个软件堆栈问题。包括：存储、超参搜索、差分渲染和光线投射等。

开放问题和挑战

- 无缝集成
- 规模化
- 通用化
- 多模态学习
- 质量

社会影响

受新神经表示法影响最大的领域是计算机视觉、计算机图形学以及增强和虚拟现实，这些领域可以从渲染环境的增强照片真实感中获益。事实上，最先进的体模型依赖于易于理解和优雅的原理，降低了摄影测量和三维重建研究的障碍。更重要的是，这些方法和公开可用的代码库和数据集的易用性放大了这种效果。

由于神经渲染还不成熟，也没有很好的理解，像Blender这样的终端用户工具还不存在，目前都无法使用这些新方法。然而，对技术的更广泛理解不可避免地会影响已开发的产品和应用。预计游戏内容创作和电影特效的工作量将减少。与现有技术相比，从几个输入图像渲染场景的照片逼真度新视图，这种可能性是一个显著的优势。这可能会重塑视觉效果（VFX）行业中的内容设计整个既定流程。

结束语

在过去几年中，神经渲染领域取得了快速发展，并继续快速增长。它的应用范围从刚性和非刚性场景的任意视点视频到形状和材质编辑、重照明和人类avatar生成等。

相信神经渲染仍然是一个新兴领域，有许多开放的挑战可以解决。