NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

NeRF：将场景表示为用于视图合成的神经辐射场

作者----

译者--Emily

摘要

我们提出了一种通过利用稀疏的输入视图来优化一个底层连续体积场景函数的方法，实现了复杂场景新视图合成的最先进效果。我们的算法使用一个全连接（非卷积）深度网络来表示场景，该网络的输入是一个连续的五维坐标（空间位置(x,y,z) 和观察方向(θ,ϕ)），输出则是在该空间位置的体积密度和视角相关的发射辐射度（emitted radiance）。我们通过沿相机光线查询五维坐标来合成视图，并使用传统的体积渲染技术将输出的颜色和密度投影到图像中。由于体积渲染本身具有可微性，所以优化我们表示所需的唯一输入是一组具有已知相机位姿的图像。在本文中，我们描述了如何通过有效地优化神经辐射场来渲染具有复杂几何和外观的场景的真实感新视图，并展示了该方法在神经渲染和视图合成方面优于以往工作的效果图。视图合成结果最好以视频形式观看，因此我们强烈建议读者查看我们的补充视频，以获得可信的对比效果。

关键字

场景表示、视图合成、基于图像的渲染、体积渲染、三维深度学习

1. 绪论

在这项工作中，我们以一种全新的方式解决了长期存在的视图合成问题，即通过直接优化连续 5D 场景表示的参数，以最小化渲染一组捕获图像的误差。

我们将静态场景表示为一个连续的 5D 函数，该函数在空间中每个点(x,y,z) 输出每个方向(θ,ϕ)发射的辐射，以及每个点的密度，密度的作用类似于微分不透明度，用于控制光线通过(x,y,z)时积累的辐射量。我们的方法通过从单个 5D 坐标(x,y,z,θ,ϕ)回归到单个体积密度和视角相关的 RGB 颜色，优化一个没有任何卷积层的深度全连接神经网络（通常称为多层感知器或 MLP）来表示该函数。

渲染这个神经辐射场（NeRF），我们提出了一种方法，可以从一组输入图像优化场景的连续五维神经辐射场表示（在任意连续位置的体积密度和视角依赖颜色）。我们使用体积渲染技术沿光线积累场景表示的样本，从而能够从任何视角渲染场景。在这里，我们展示了在环绕半球上随机捕捉的合成鼓场景的 100 个输入视图集合，并展示了从我们优

化的 NeRF 表示渲染的两个新视图。

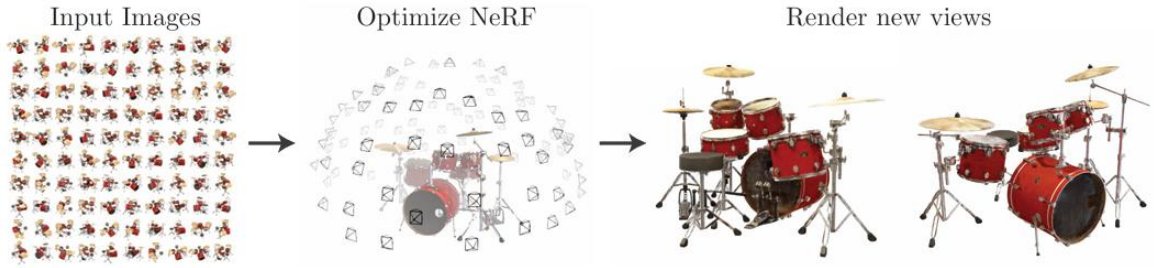

图 1

从特定的视角来看，我们执行以下操作：

1）沿着场景对摄像机光线进行行进，以生成一个采样的 3D 点集合；

2）使用这些点及其对应的二维观察方向作为神经网络的输入，生成输出的颜色和密度集合；

3）使用经典的体积渲染技术将这些颜色和密度累积成二维图像。由于该过程天然可微分，我们可以使用梯度下降来通过最小化每个观测图像与从我们的表示渲染的对应视图之间的误差来优化该模型。在多个视图上最小化该误差会促使网络通过为包含真实场景内容的位置分配高体密度和准确颜色来预测一致的场景模型。图 2 展示了这一整体流程。

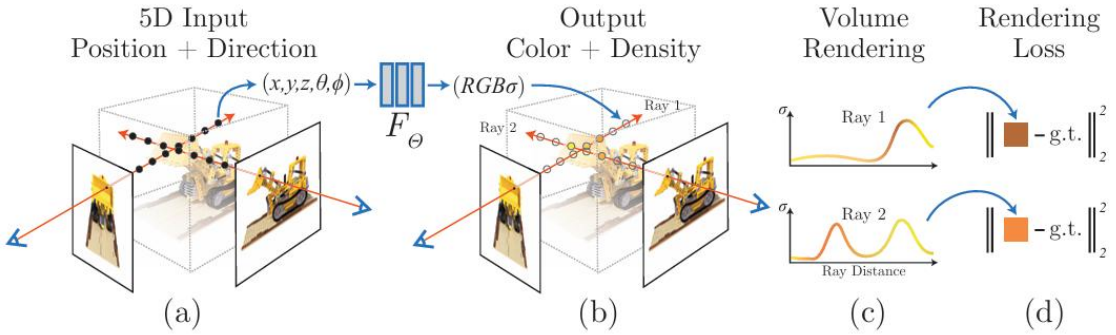

图 2

从图 2 看到，我们神经辐射场场景表示和可微渲染过程的概述。我们通过在相机光线沿线采样 5D 坐标（位置和观察方向）（a），将这些位置输入多层感知机（MLP）以生成颜色和体积密度（b），并使用体积渲染技术将这些值合成为图像（c）。该渲染函数是可微的，因此我们可以通过最小化合成图像与真实观测图像之间的残差来优化我们的场景表示（d）。

我们发现，对于复杂场景优化神经辐射场表示的基本实现无法收敛到足够高分辨率的表示，并且在每条相机光线所需样本数量上效率低下。我们通过使用位置编码对输入的 5D 坐标进行变换来解决这些问题，使得多层感知机（MLP）能够表示更高频的函数；同时，我们提出了一种分层采样程序，以减少充分采样这种高频场景表示所需的查询次数。我们的方法继承了体积表示的优势：两者都可以表示复杂的真实世界几何和外观，并且非常适合使用投影图像进行基于梯度的优化。关键在于，我们的方法克服了在高分辨率下建模复杂场景时离散体素网格所带来的高昂存储成本。总之，我们的技术贡献包括：

- 一种将具有复杂几何形状和材质的连续场景表示为五维神经辐射场的方法，该方法使用基础的多层感知器（MLP）网络参数化。

- 一种基于经典体积渲染技术的可微渲染过程，我们利用该过程从标准 RGB 图像中优化这些表示。这包括一种分层采样策略，用于将 MLP 的容量分配到具有可见场景内容的空间中。
- 一种位置编码方法，将每个输入的五维坐标映射到更高维的空间，从而使我们能够成功优化神经辐射场以表示高频场景内容。

我们展示了我们所提出的神经辐射场方法在量化和质化上都优于最先进的视图合成方法，包括将神经三维表示拟合到场景的工作以及训练深度卷积网络以预测采样体积表示的工作。据我们所知，本文提出了首个连续神经场景表示，能够基于在自然环境中捕获的 RGB 图像渲染出真实物体和场景的高分辨率逼真新视图。

2. 相关工作

计算机视觉中一个有前途的最新方向是将对象和场景编码到多层感知机（MLP）的权重中，该 MLP 可以直接从三维空间位置映射到形状的隐式表示，例如该位置的符号距离 [6]。然而，到目前为止，这些方法还无法像使用离散表示（如三角网格或体素网格）表示场景的技术那样以同样的精度重现复杂几何的逼真场景。在本节中，我们将回顾这两类工作，并将其与我们的方法进行对比，我们的方法增强了神经场景表示的能力，以生成复杂逼真场景渲染的最先进结果。使用 MLP 将低维坐标映射到颜色的类似方法也已用于表示其他图形函数，例如图像 [44]、纹理材料 [12,31,36,37] 以及间接光照值 [38]。

神经 3D 形状表示

近期的研究调查了将连续 3D 形状隐式表示为水平集的方法，通过优化深度网络将 xyz 坐标映射到有符号距离函数 [15,32] 或占据场 [11,27]。然而，这些模型的局限在于它们需要访问真实的 3D 几何数据，这通常来源于诸如 ShapeNet [3] 的合成 3D 形状数据集。随后的研究通过制定可微分的渲染函数，放宽了对真实 3D 形状的要求，使神经隐式形状表示仅使用 2D 图像就能进行优化。Niemeyer 等 [29] 将表面表示为 3D 占据场，并使用数值方法找到每条射线的表面交点，然后使用隐式微分计算精确导数。每个射线交点位置作为输入提供给神经 3D 纹理场，该纹理场预测该点的漫反射颜色。Sitzmann 等 [42] 使用一种不太直接的神经 3D 表示，在每个连续 3D 坐标输出特征向量和 RGB 颜色，并提出了一种可微分渲染函数，该函数由一个循环神经网络组成，沿每条射线推进以决定表面位置。

虽然这些技术有可能表示复杂和高分辨率的几何形状，但到目前为止，它们仅限于几何复杂度较低的简单形状，导致渲染结果过于平滑。我们展示了一种替代策略，即优化网络以编码五维辐射场（3D 体积与 2D 视图相关外观），可以表示更高分辨率的几何形状和外观，从而渲染复杂场景的真实感新视角。

视图合成与基于图像的渲染

给定密集的视图采样，通过简单的光场采样插值技术 [21,5,7] 可以重建逼真的新视图。对于稀疏视图采样的新视图合成，计算机视觉和图形学社区通过从观测图像中预测传统的几何和外观表示取得了显著进展。一类流行的方法使用基于网格的场景表示，其外观可以是漫反射的 [48] 或视角相关的 [2,8,49]。可微分光栅化器 [4,10,23,25] 或路径追踪器 [22,30] 可以直接优化网格表示，以使用梯度下降重现一组输入图像。然而，基于图像重投影的梯度网格优化通常很困难，这可能是由于局部极小值或损失函数景观的条件不良。此外，这种策略需要在优化之前提供具有固定拓扑的模板网格 [22] 作为初始化，而在不受约束的真实世界场景中，这通常是不可用的。

另一类方法使用体积表示来处理从一组输入 RGB 图像生成高质量逼真视图的任务。体积方法能够真实地表示复杂形状和材料，非常适合基于梯度的优化，并且相比基于网格的方法，产生的视觉干扰伪影通常更少。早期的体积方法使用观测到的图像直接对体素网格进行上色 [19,40,45]。最近，一些方法 [9,13,17,28,33,43,46,52] 使用多个场景的大型数据集训练深度网络，从一组输入图像预测采样的体积表示，然后在测试时使用 alpha 合成 [34] 或沿光线学习的合成来渲染新视图。其他工作则针对每个特定场景优化卷积网络（CNN）与采样体素网格的组合，使 CNN 能够补偿低分辨率体素网格的离散化伪影 [41]，或允许根据输入时间或动画控制变动预测的体素网格 [24]。尽管这些体积技术在新视图合成方面取得了令人印象深刻的成果，但由于离散采样导致的时间和空间复杂度较高，其扩展到更高分辨率图像的能力受到根本限制——渲染高分辨率图像需要对三维空间进行更细的采样。我们通过改为在深度全连接神经网络的参数中编码连续体积来规避这一问题，这不仅比以往的体积方法生成显著更高质量的渲染图像，而且存储成本仅为那些采样体积表示的一小部分。

3. 神经辐射场场景表示

我们将连续场景表示为一个五维向量值函数，其输入为三维位置 x = (x y z) 和二维视角方向(θ,ϕ)，输出为发射颜色 c = (r g b) 和体积密度。

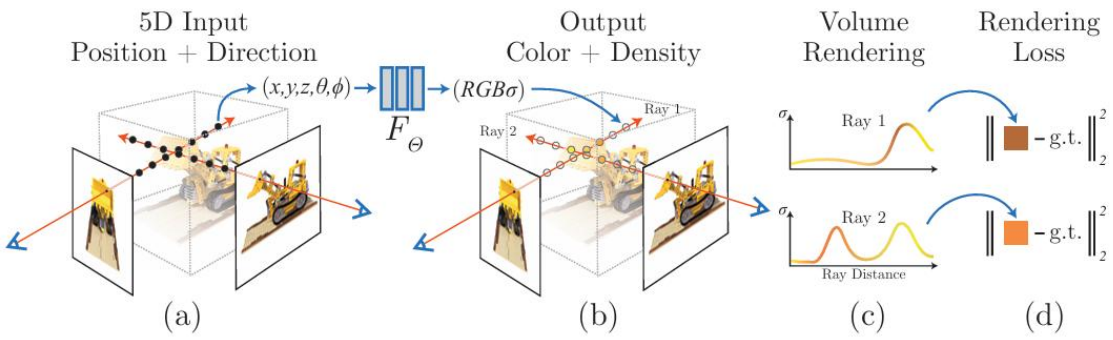

图 2

从图 2 看到，我们神经辐射场场景表示和可微渲染过程的概述。我们通过在相机光线沿线采样 5D 坐标（位置和观察方向）（a），将这些位置输入多层感知机（MLP）以生成

颜色和体积密度（b），并使用体积渲染技术将这些值合成为图像（c）。该渲染函数是可微的，因此我们可以通过最小化合成图像与真实观测图像之间的残差来优化我们的场景表示（d）。

我们将方向表示为三维笛卡尔单位向量 d。我们使用一个多层感知机（MLP）网络 F_Θ:(*x*,*d*)−>(c, σ)来近似这种连续的五维场景表示，并优化其权重，使其能够将每个输入的五维坐标映射到相应的体积密度和方向发射颜色。

我们通过限制网络仅将体积密度预测为位置 x 的函数，同时允许 RGB 颜色 c 作为位置和视角方向的函数进行预测，从而鼓励表示在多视图下保持一致。为此，MLP F_Θ首先使用 8 个全连接层（每层 256 个通道，ReLU(激活函数) 激活）处理输入的 3D 坐标 x，并输出一个 256 维的特征向量。然后将该特征向量与相机光线的观察方向拼接，并传递给另外一个全连接层（使用 ReLU(激活函数) 激活，128 个通道），输出依赖视角的 RGB 颜色。

请参见图 3，了解我们的方法如何使用输入的观察方向来表示非朗伯效应。如图 4 所示，在没有视角依赖（仅以 x 为输入）的情况下训练的模型在表示高光时存在困难。

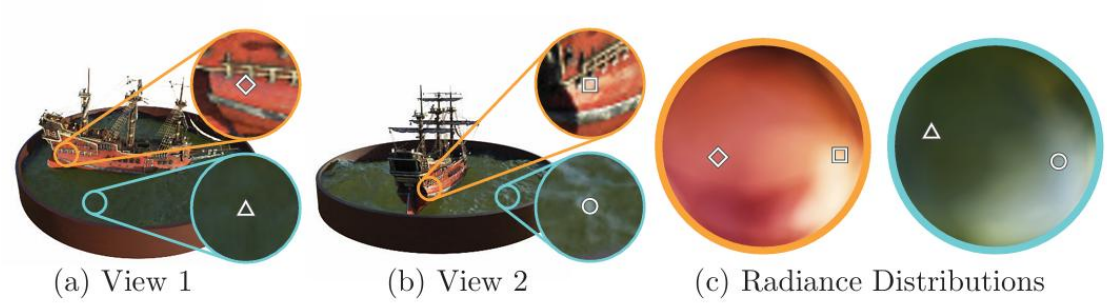

图 3

从图 3 来看，视角依赖发射辐射的可视化。我们的神经辐射场表示将 RGB 颜色输出为空间位置 x 和观察方向 d 的五维函数。在此，我们展示了船舶场景中神经表示的两个空间位置的示例方向颜色分布。在 (a) 和 (b) 中，我们展示了从两个不同相机位置观察两个固定 3D 点的外观：一个在船舶侧面（橙色插图），一个在水面上（蓝色插图）。我们的方法可以预测这两个 3D 点随视角变化的高光外观，并且在 (c) 中，我们展示了这种行为如何在整个观察方向半球上连续推广。

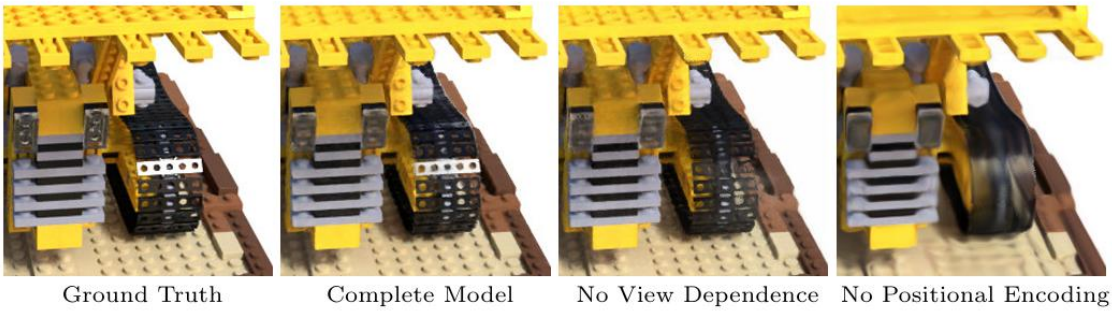

图 4

从图 4 来看，在这里，我们展示了我们的完整模型如何从表示与视角相关的发射辐射以及通过高频位置编码处理输入坐标中受益。去除视角依赖会使模型无法重现推土机履带

上的镜面反射。去除位置编码会大幅降低模型表示高频几何和纹理的能力，从而会导致过度平滑的外观。

4. 使用辐射场的体积渲染

我们的 5D 神经辐射场将一个场景表示为空间中任意点的体积密度和方向发射辐射。我们使用经典体积渲染的原理[16]来渲染通过场景的任意光线的颜色。体积密度σ(x)可以理解为光线在位置 (x) 停止于微小粒子的微分概率。相机光线 r(t)=o+td 在近远界t_n和 t_f 之间的期望颜色 C(r)为：

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right). \quad (1)$$

函数 T(t)表示沿射线从t_n到 t 的累积透射率，即射线从t_n到 t 在没有碰到任何其他粒子的情况下传播的概率。从我们的连续神经辐射场渲染视图需要对每个虚拟相机像素追踪的相机射线来估计此积分 C(r)。我们使用求积法数值估计该连续积分。确定性求积法通常用于渲染离散化的体素网格，但会有效地限制我们的表示分辨率，因为 MLP 只会在固定的离散位置被查询。相反，我们使用分层采样方法，将$[t_n,t_f]$分成 N 个等间距的区间，然后在每个区间内随机均匀抽取一个样本：

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n),\ t_n + \frac{i}{N}(t_f - t_n)\right]. \quad (2)$$

虽然我们使用离散样本集来估计积分，但分层采样使我们能够表示连续的场景表示，因为在优化过程中，MLP 会在连续的位置上进行评估。我们使用这些样本根据 Max [26]的体渲染综述中讨论的求积规则来估计 C(r) ：

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i\,, \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1}\sigma_j\delta_j\right), \quad (3)$$

其中$\delta_i = t_{i+1} - t_i$ 是相邻样本之间的距离。用于从(c_i,σ_i) 值集合计算 $\hat{C}$(r)的函数可以轻松求导，并且在 $\alpha_i = 1 - e^{-\sigma_i\delta_i}$ 时简化为传统的 α 合成。

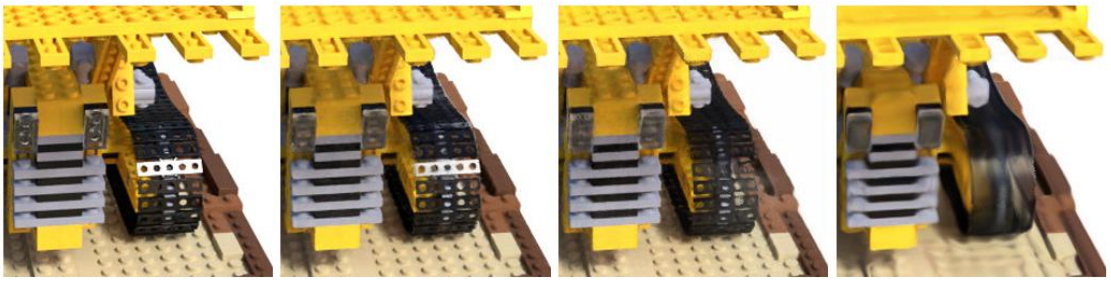

图 4

从图 4 来看，在这里，我们展示了我们的完整模型如何从表示与视角相关的发射辐射以及通过高频位置编码处理输入坐标中受益。去除视角依赖会使模型无法重现推土机履带上的镜面反射。去除位置编码会大幅降低模型表示高频几何和纹理的能力，导致过度平滑

的外观。

5. 优化神经辐射场

在前一节中，我们描述了将场景建模为神经辐射场并从该表示生成新视图所需的核心组件。然而，我们观察到这些组件不足以实现最先进的质量，如第 6.4 节所示。我们引入了两项改进，以便能够表示高分辨率复杂场景。第一项是对输入坐标进行位置编码，以帮助多层感知机（MLP）表示高频函数；第二项是分层采样过程，使我们能够高效地采样这种高频表示。

5.1 位置编码

尽管神经网络是通用函数近似器 [14]，我们发现让网络F_θ直接操作(xyzθϕ)输入坐标会导致渲染效果在表示颜色和几何的高频变化时表现不佳。这与 Rahaman 等人 [35] 最近的研究一致，该研究表明深度网络倾向于学习低频函数。他们还显示，在将输入映射到使用高频函数的高维空间后再传递给网络，可以更好地拟合包含高频变化的数据。我们在神经场景表示的背景下利用这些发现，并展示将F_θ重新表述为两个函数的组合$F_\Theta = F'_\Theta \circ \gamma$，其中一个是学习得到的，另一个不是，可以显著提高性能（参见图 4 和表 2）。这里是从 R 映射到高维空间 R^{2L}的映射，而F_Θ 仍然只是一个普通的 MLP。通常，我们使用的编码函数是：

$$\gamma(p) = \left(\sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right), \cdots, \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\right). \qquad (4)$$

这个函数γ(·) 被分别应用于位置 x 中的三个坐标值（它们被标准化以位于 [1,1] 范围内）以及笛卡尔视角单位向量 d 的三个分量（根据构造，它们位于 [1,1] 范围内）。在我们的实验中，我们为γ(x)设置 L = 10，为γ(d)设置 L = 4。在流行的 Transformer 架构 [47] 中也使用了类似的映射，这里被称为位置编码。然而，Transformer 使用它的目的是为不包含任何顺序概念的架构提供序列中标记的离散位置作为输入。相比之下，我们使用这些函数将连续的输入坐标映射到更高维的空间，以使我们的 MLP 更容易逼近高频函数。在从投影建模 3D 蛋白质结构的相关问题的同步研究中 [51]，也使用了类似的输入坐标映射。

5.2 分层体积采样

通常情况下，沿每条相机射线在 N 个查询点密集评估神经辐射场网络的渲染策略效率低下：那些不对渲染图像做出贡献的空旷区域和遮挡区域仍然会被重复采样。我们从早期的体积渲染工作中汲取灵感[20]，提出了一种分层表示方法，通过按对最终渲染的预期影响分配采样点，从而提高渲染效率。

我们并不是仅使用单个网络来表示场景，而是同时优化两个网络：一个粗略的网络和一个精细的网络。我们首先使用分层采样（stratified sampling）对N_c个位置进行采样，并

按照公式 2 和 3 对这些位置上的粗略网络进行评估。根据这个粗略网络的输出，我们随后在每条射线沿线生成更为合理的采样点，其中样本偏向于体积的相关部分。为此，我们首先将公式 3 中粗略网络的 alpha 复合颜色 $\hat{C}_c(r)$重写为沿射线的所有采样颜色c_i 的加权和：

$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i \,, \qquad w_i = T_i(1 - \exp(-\sigma_i \delta_i)) \,. \tag{5}$$

将这些权重归一化为 $\hat{w}_i=\frac{w_i}{\sum_{j=1}^{N_c} w_j}$ 会沿光线产生分段常数的概率密度函数（PDF）。

我们使用逆变换采样从该分布中采样第二组N_f个位置，在第一组和第二组样本的合集上评估我们的神经网络，并使用公式 3 但利用所有N_c+N_f个样本计算光线的最终渲染颜色 $\hat{C}_c(r)$。该方法会将更多样本分配到我们预计包含可见内容的区域。这实现了与重要性采样类似的目标，但我们将采样值作为整个积分域的非均匀离散化，而不是将每个样本视为整个积分的独立概率估计。

5.3 实现细节

我们为每个场景优化一个独立的神经连续体积表示网络。这只需要一个场景的捕获 RGB 图像数据集、对应的相机位姿和内参，以及场景边界（对于合成数据，我们使用真实的相机位姿、内参和边界；对于实际数据，我们使用 COLMAP 的结构光束法（structure-from-motion）软件包 [39] 来估计这些参数）。在每次优化迭代中，我们从数据集中所有像素的集合中随机采样一批相机光线，然后按照第 5.2 节描述的分层采样从粗网络查询N_c个样本，从细网络查询N_c+N_f个样本。接着，我们使用第 4 节描述的体渲染过程来渲染两组样本中每条光线的颜色。我们的损失函数就是粗渲染和细渲染的渲染像素颜色与真实像素颜色之间的总平方误差。

$$\mathcal{L} = \sum_{\mathbf{r}\in\mathcal{R}} \left[\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right] \tag{6}$$

其中 R 是每个批次中的光线集合，$C(r)$、$\hat{C}_c(r)$ 和 $\hat{C}_f(r)$分别表示光线 r 的真实值、粗略体积预测的 RGB 颜色和精细体积预测的 RGB 颜色。请注意，尽管最终渲染结果来自$\hat{C}_f(r)$，我们仍会最小化$\hat{C}_c(r)$的损失，以便粗略网络的权重分布可以用于在精细网络中分配采样。在我们的实验中，我们使用批量大小为 4096 条光线，每条光线在粗略体积中采样N_c= 64 个坐标，在精细体积中额外采样N_f= 128 个坐标。我们使用 Adam 优化器 [18]，初始学习率为 5×10^{-4}，并在优化过程中指数衰减到 5×10^{-5}（其他 Adam 超参数保持默认值，即 $\beta1 = 0.9, \beta2 = 0.999, \varepsilon = 10^{-7}$）。单个场景的优化通常需要约 100k 到 300k 次迭代在单个 NVIDIA V100 GPU 上收敛（大约 12 天）。

6. 结论

我们通过定量分析（表 1）和定性分析（图 8 和图 6）展示了我们的方法优于以往的工作，并提供了大量消融实验以验证我们的设计选择（表 2）。我们建议读者观看我们的补充视频，以更好地理解我们的方法在渲染新视角的平滑路径时，相较于基线方法所取得的显著改进。

6.1 数据集

物体的合成渲染

我们首先展示了两个物体合成渲染数据集的实验结果（表 1，Diffuse Synthetic 360 和 Realistic Synthetic 360）。DeepVoxels [41] 数据集包含四个具有简单几何形状的朗伯体物体。每个物体都以 512×512 像素呈现，从上半球采样的视点进行渲染（输入为 479 个视点，测试为 1000 个视点）。此外，我们还生成了自己的数据集，其中包含八个具有复杂几何形状和真实非朗伯材质的物体的路径追踪图像。六个物体从上半球采样的视点渲染，两个位物体从全球采样的视点渲染。我们为每个场景渲染了 100 个视角作为输入，测试用渲染 200 个视角，所有分辨率均为 800×800 像素。

	Diffuse Synthetic 360° [41]			Realistic Synthetic 360°			Real Forward-Facing [28]		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SRN [42]	33.20	0.963	0.073	22.26	0.846	0.170	22.84	0.668	0.378
NV [24]	29.62	0.929	0.099	26.05	0.893	0.160	-	-	-
LLFF [28]	34.38	0.985	0.048	24.88	0.911	0.114	24.13	0.798	**0.212**
Ours	**40.15**	**0.991**	**0.023**	**31.01**	**0.947**	**0.081**	**26.50**	**0.811**	0.250

表 1：我们的方法在合成图像和真实图像的数据集上定量地优于以往工作。我们报告了 PSNR/SSIM（越高越好）和 LPIPS [50]（越低越好）。DeepVoxels [41] 数据集由 4 个具有简单几何形状的漫射物体组成。我们的真实感合成数据集包含 8 个具有复杂几何形状和复杂非朗伯材质的对象的路径追踪渲染。真实数据集由手持设备拍摄的 8 个现实场景的前向视图组成（NV 无法在此数据上评估，因为它只能重建有限体积内的物体）。尽管 LLFF 的 LPIPS 略好，我们仍建议读者观看我们的补充视频，在视频中我们的方法实现了更好的多视图一致性，并比所有基线产生更少的伪影。

复杂场景的真实图像

我们展示了在用大致前置拍摄的图像捕捉的复杂真实世界场景上的结果（表 1，真实前置）。该数据集包含 8 个场景，使用手持手机拍摄（其中 5 个来自 LLFF 论文，3 个是我们拍摄的），拍摄时每个场景包含 20 到 62 张图像，并将其中 18 张保留作为测试集。所有图像的分辨率为 1008 × 756 像素。

6.2 比较

为了评估我们的模型，我们将其与当前在视图合成方面性能最好的技术进行比较，

具体如下。除了 Local Light Field Fusion [28] 外，所有方法都使用相同的输入视图集为每个场景训练一个独立的网络，而 Local Light Field Fusion [28] 则是在一个大型数据集上训练一个单一的 3D 卷积网络，然后在测试阶段使用相同的训练网络处理新场景的输入图像。

神经体积（Neural Volumes，NV）[24]

合成位于前景有限体积内的物体的新视角（背景必须单独拍摄，不包含目标物体）。它优化一个深度 3D 卷积网络，以预测一个分辨率为 128^3 的离散 RGB 体素网格以及 323 采样的 3D 变形网格。该算法通过沿着变形体素网格遍历相机光线来渲染新视角。

场景表示网络（Scene Representation Networks, SRN）[42]

它将连续场景表示为不透明表面，通过一个多层感知机（MLP）隐式定义，该 MLP 将每个 (x, y, z) 坐标映射到特征向量。他们训练一个循环神经网络沿着光线遍历场景表示，使用任何 3D 坐标的特征向量来预测沿光线的下一步距离。最终步的特征向量会被解码为表面上该点的单一颜色。需要注意的是，SRN 是同一作者对 DeepVoxels [41] 的性能改进版本，这也是我们没有将 DeepVoxels 纳入比较的原因。

图 5

从图 5 看出，对来自我们使用基于物理的渲染器生成的新合成数据集场景的测试集视图进行比较。我们的方法能够在几何和外观上恢复精细细节，例如船只的索具、乐高的齿轮和履带、麦克风的光亮支架和网格罩，以及材料的非朗伯反射（non Lambertian refectance）。LLFF 在麦克风支架和材料物体边缘出现条纹伪影，在船只桅杆和乐高物体内部出现重影

伪影。SRN 在所有情况下产生模糊和失真的渲染结果。Neural Volumes 无法捕捉麦克风网格罩或乐高齿轮的细节，并且完全无法恢复船只索具的几何结构。

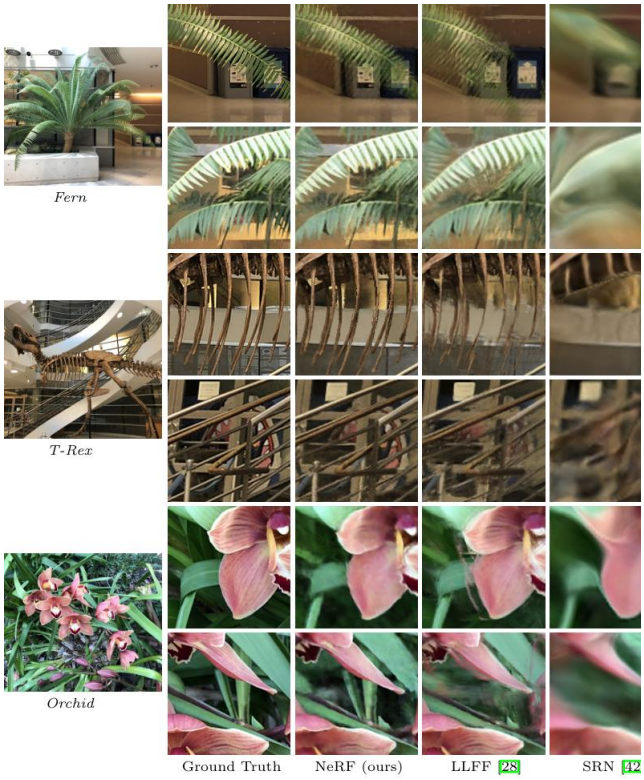

图 6

从图 6 看出，对现实世界场景的测试集视图进行比较。LLFF 专门为此使用场景设计（面向前方的真实场景捕捉）。如 Fern 的叶子以及 T-rex 的骨架肋骨和栏杆所示，我们的方法能够比 LLFF 更一致地在渲染视图之间表示细微几何结构。我们的方法还能够正确重建 LLFF 难以干净渲染的部分遮挡区域，例如底部 Fern 裁剪中叶子后面的黄色货架，以及底部 Orchid 裁剪背景中的绿色叶子。多重渲染之间的混合也可能导致 LLFF 中出现重复边缘，如顶部 Orchid 裁剪所示。SRN 捕捉了每个场景中的低频几何和颜色变化，但无法再现任何细节。

6.3 讨论

我们在所有场景中都彻底超越了两个基线方法，这两个方法同样为每个场景优化一个独立的网络（NV 和 SRN）。此外，我们在使用仅其输入图像作为整个训练集的情况下，所生成的渲染在质量和数量上都优于 LLFF（除一个指标外）。SRN 方法会产生高度平滑的几何和纹理，它在视图合成方面的表现能力受限于每条相机光线只选择单一深度和颜色。NV 基线能够捕捉到相当详细的体积几何和外观，但其使用底层的显式 128^3 体素网格限制了其在高分辨率下表示细节的能力。LLFF 特别提供了采样指导，要求输入视图之间的视差不超过 64 像素，因此在包含高达 400-500 像素视差的合成数据集中，LLFF 经常无法正确估计几何。此外，LLFF 在渲染不同视图时，会在不同场景表示之间进行混合，导致感知上明显的不一致，这在我们的补充视频中很明显。实用上，这些方法之间最大的权衡在于时间与空间。所有比较的单场景方法每个场景至少需要 12 小时的训练时间。相比之下，LLFF 可以在不到 10 分钟内处理一个小型输入数据集。然而，LLFF 为每张

输入图像生成一个大的 3D 体素网格，从而导致巨大的存储需求（一个真实合成场景超过 15GB）。我们的方法仅需要 5 MB 来存储网络权重（相较于 LLFF 压缩比为 3000），甚至比单场景的数据集输入图像所需的内存还少。

6.4 消融研究

我们通过表 2 中的广泛消融研究验证了算法的设计选择和参数。我们展示了在现实感合成 360 场景上的结果。第 9 行显示了我们的完整模型，作为参考点。第 1 行显示了一个精简版本的模型，没有位置编码 (PE)、视图依赖性 (VD) 或分层采样 (H)。在第 2 至第 4 行中，我们从完整模型中一次移除这三个组件，观察到位置编码（第 2 行）和视图依赖性（第 3 行）提供了最大的定量收益，其次是分层采样（第 4 行）。第 5 至第 6 行显示了随着输入图像数量减少，我们的性能如何下降。需要注意的是，即使仅使用 25 张输入图像，我们的方法在所有指标上仍超过 NV、SRN 和 LLFF，即使它们使用了 100 张图像（参见补充材料）。在第 7 至第 8 行中，我们验证了最大频率的选择。

	Input	#Im.	L	(N_c, N_f)	PSNR↑	SSIM↑	LPIPS↓
1) No PE, VD, H	xyz	100	-	(256, -)	26.67	0.906	0.136
2) No Pos. Encoding	$xyz\theta\phi$	100	-	(64, 128)	28.77	0.924	0.108
3) No View Dependence	xyz	100	10	(64, 128)	27.66	0.925	0.117
4) No Hierarchical	$xyz\theta\phi$	100	10	(256, -)	30.06	0.938	0.109
5) Far Fewer Images	$xyz\theta\phi$	25	10	(64, 128)	27.78	0.925	0.107
6) Fewer Images	$xyz\theta\phi$	50	10	(64, 128)	29.79	0.940	0.096
7) Fewer Frequencies	$xyz\theta\phi$	100	5	(64, 128)	30.59	0.944	0.088
8) More Frequencies	$xyz\theta\phi$	100	15	(64, 128)	30.81	0.946	0.096
9) Complete Model	$xyz\theta\phi$	100	10	(64, 128)	**31.01**	**0.947**	**0.081**

表 2：我们模型的消融研究。指标在我们真实感合成数据集的 8 个场景上取平均。详细说明见第 6.4 节。

L 用于我们对 x 的位置编码（用于 d 的最大频率按比例缩放）。仅使用 5 个频率会降低性能，但将频率数量从 10 增加到 15 并不会提升性能。我们认为，一旦 2L 超过采样输入图像中存在的最大频率（在我们的数据中大约为 1024），增加 L 的好处是有限的。

7. 总结

我们的工作直接针对以往使用多层感知器（MLP）将物体和场景表示为连续函数的研究存在的不足。我们展示了将场景表示为 5D 神经辐射场（一个 MLP，根据三维位置和二维视角输出体积密度和视角依赖的发射辐射）比以往占主导地位的训练深度卷积网络输出离散体素表示的方法产生更好的渲染效果。尽管我们提出了一种分层采样策略以提高渲染的采样效率（用于训练和测试），在研究如何高效优化和渲染神经辐射场方面仍有很大进展空间。未来工作的另一个方向是可解释性：像体素网格和网格这样的采样表示可以推理渲染视图的预期质量和失败模式，但当我们将场景编码在深度神经网络的权重中时，这些

问题的分析方法尚不明确。我们相信，这项工作为基于真实世界图像的图形流水线取得了进展，其中复杂场景可以由从实际物体和场景图像优化得到的神经辐射场构成。

References

略

附录 A---附加实现细节

略

附录 B---附加基准方法详情

略

附录 C---NDC 射线空间导出

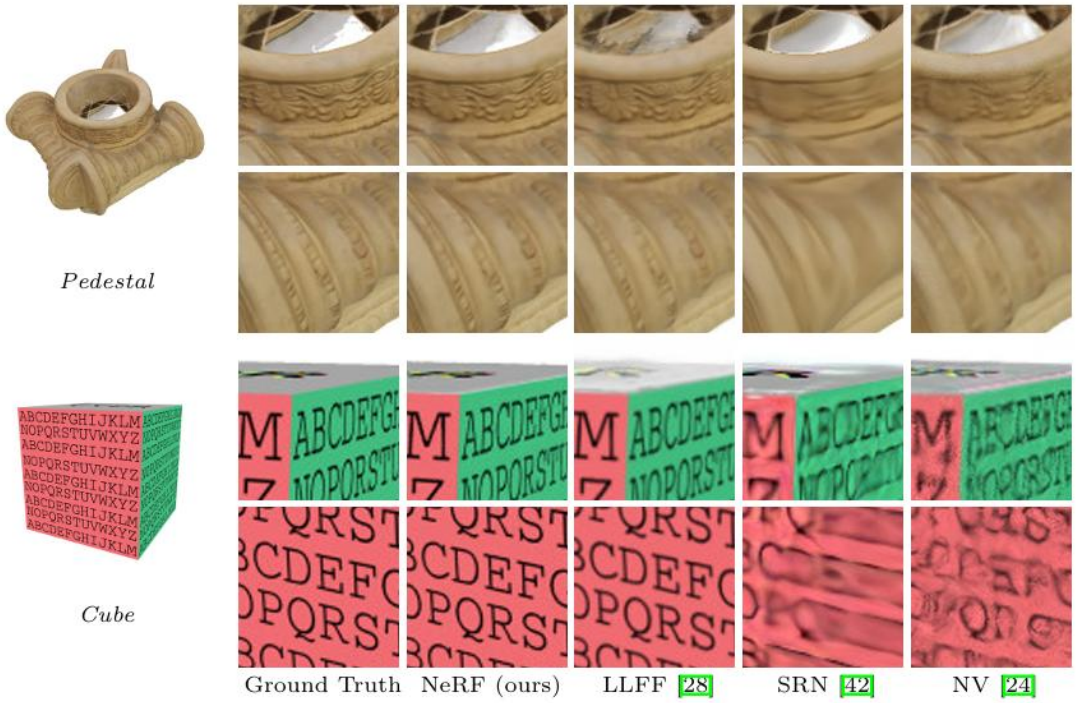

图 8

从图 8 可看出，来自 DeepVoxels [41] 合成数据集场景的测试集视图比较。该数据集中的物体具有简单的几何形状和完美的漫反射特性。由于输入图像数量众多（479 个视图）

且渲染物体简单，我们的方法和 LLFF [28] 在该数据上几乎都能完美表现。LLFF 在对其 3D 体积进行插值时仍偶尔会出现伪影，如每个物体顶部的插图所示。SRN [42] 和 NV [24] 在渲染细节方面的表示能力不足。

附录 D---更多结果

略