融合不确定性估计和 NeRF 的隐式三维建模方法设计与实现

B20031922 周帅 [1]

[1] Nanjing University of Posts and Telecommunications

摘要

随着计算机视觉和图形学领域的不断发展，三维场景的建模和重建成为近年来备受关注的技术焦点，特别是在虚拟现实（VR）和增强显示（AR）等领域，对于精确、高效的三维场景重建需求日益增长。目前主流的三维重建方法有显示和隐式两种方式，但均存在一些局限和缺点：显式三维重建方法占用空间大、存储利用率低；隐式三维重建学习能力有限，难以重建场景细节。相比之下，NeRF（神经辐射场）作为一种新兴的隐式三维重建方法，具有独特的优势。它利用神经网络对场景中的三维对象进行建模，学习出场景信息，并结合体渲染技术合成高质量的新视角视图，克服了传统方法在处理复杂场景时面临的局限性。本文实现了融合不确定性估计和 NeRF 的隐式三维建模方法。该算法以 NeRF 为基础，通过引入不确定性估计，对模型预测的不确定性进行有效评估，为重建场景中数据缺失或模糊的区域提供置信度信息。此外，融合不确定性估计的 NeRF 不仅提升了模型在少量观察样本下的鲁棒性，而且为理解场景提供了更深入的视角，同时以最小的额外资源消耗提升了新视角合成的质量。并且引入主动学习的策略，使用不确定性估计评估新输入的效果，以最有效的方式补充训练样本。实验结果显示，该算法在真实场景和合成场景中的性能表现较好，尤其是在训练数据较为有限的情况下，其优势更加明显。

关键词：NeRF；三维建模；不确定性估计；神经网络

1. 引言

在数字时代的浪潮中，三维场景的建模和重建技术已经跃升为多个领域不可或缺的核心驱动力。随着科技的飞速发展，人们不再满足于二维平面的展示和交互，而是追求更加真实、立体的三维体验。因此，三维建模和重建技术应运而生，并迅速在游戏开发、虚拟现实、影视制作、建筑设计、医学成像等众多领域展现出其独特的价值。如图1，展示了三维重建技术在文物建筑重建的具体应用早期的手工三维场景建模使用（Computer-Aided Design，CAD）软件或数字内容创作（Digital Content Creation，DCC）软件进行操作和创建。但是使用 CAD 软件或 DCC 软件进行手工三维建模需要一定的技术和经验。用户通常需要了解软件的操作界面、建模工具和技术，以及三维建模的基本原理和规范。他们还需要根据具体的设计需求和场景要求，调整和优化建模结果，以达到所需的效果。但它们通常需要大量的人力和时间投入，并且对操作者的技能和经验要求较高。在计算机中对三维场景进行数字化表示和构建是现今诸多重要应用的基础。与以往依赖软件或数字内容创作工具进行手工三维建模的方式有所区别，计算机中三维重建技术的目标是通过传感器输入，如图片、三维点云等数据，自动重新构建出相应的三维结构和场景，全程无需人工介入。目前主流的三维重建方法有显示和隐式两种方式，但均存在一些局限和缺点：显式三维重建是一种基于几何体的方法，其中场景的几何形状通过显式的表示方式进行建模，该方法占用空间大、存储利用率低；隐式三维重建隐式三维建模是一种基于函数表示的方法，其中场景的表面或体积是通过一个函数进行隐式定义的，该

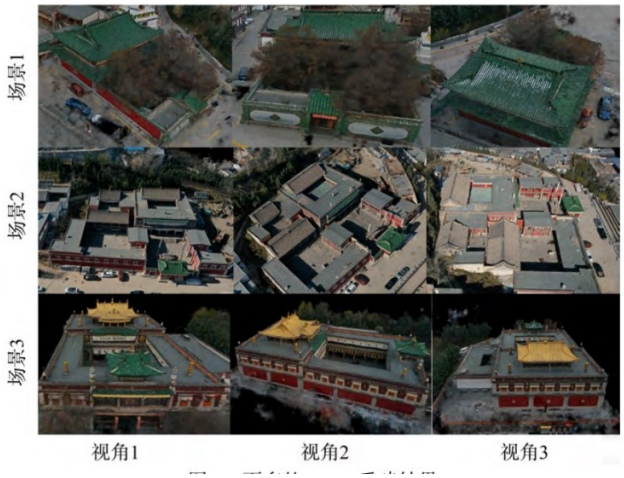

图 1. 三维建模的应用

方法学习能力有限，需要大量数据支持。随着深度学习技术的飞速进步，研究者们开始转向神经网络，以探索其在三维场景重建和图像合成领域的潜力。在这一过程中，为了突破传统体积渲染的局限性，神经辐射场（Neural Radiance Fields，NeRF）这一新兴技术应运而生，为相关领域的研究提供了全新的解决方案。它利用神经网络对场景中的三维对象进行建模，学习出场景信息，并结合体渲染技术合成高质量的新视角视图，克服了传统方法在处理复杂场景时面临的局限性。在重建 3D 场景和从稀疏的 2D 图像合成新视图方面表现出了良好的性能。尽管 NeRF 是有效的，但其性能受到训练样本质量的高度影响。由于来自场景的有限的摆拍图像，NeRF 不能很好地推广到新的视图，并且可能在未观察到的区域崩溃为平凡的解决方案。在三维重建和建模过程中，不确定性估计的作用不可忽视。由于真实场景往往包含噪声、遮挡和不完整的数据，因此对于重建结果的可靠性评估至关重要。传统的三维重建方法往往没有估计场景中各个点的不确定性，这可能导致重建结果的失真和不可靠性。通过引入不确定性估计，可以确保在观测点较少的情况下保持模型的鲁棒性，并提供对场景理解的解释。因此可以获得更准确和可靠的三维场景表示，并提供与重建结果相关的可信度度量。这将有助于改善三维重建的准确性，并提供更可靠的决策依据。基于不确定性估计，三维场景重建可以采用主动学习方案在现有训练集中补充新采样。通过评估新输入数据带来的不确定性减少程度，选择提供最大信息增益的样本。通过这种算法可以在最小程度的额外资源投入下提升新视角合成的质量。综合来看，融合不确定性估计和 NeRF 的隐式三维建模具有重要的研究意义和广阔的应用前景。它可以提高三维重建的准确性和可靠性，增强对复杂场景的表达能力，并且提供对场景的解释，以满足多个领域对高质量三维场景建模和渲染的需求。

2. 相关工作

伴随着深度学习技术的日益发展，特别是在计算机视觉与图形学的领域中，隐式三维建模方式已然成为备受瞩目的研究焦点。

2.1. 神经 3D 形状表示

传统的三维形状表示方法通常使用离散的网格或体素来表示物体的形状，然而，这种离散表示方式在捕捉细节和处理连续曲面时存在一些限制。神经 3D 形状表示是一种利用神经网络来表示三维物体或场景形状的方法。它通过神经网络学习和表示物体的几何形状、外观属性以及其他相关信息。早期的工作运用了深度神经网络，将三维空间中的坐标 x,y,z 映射为有符号距离函数或占据场。然而，这些方法的一个显著局限在于它们对真实三维几何形状数据的依赖，这些数据通常来源于合成的三维形状数据集，例如 ShapeNet[1]。这种对真实数据的直接需求极大地限制了模型在真实世界场景中的实际应用。为了突破既有的限制，后续研究采取了引入可微分渲染函数的新策略，使神经隐式形状表示能够仅凭二维图像进行高效优化。通过这一策略，神经隐式形状表示无需其他额外信息，仅依赖二维图像即可进行精细优化。Niemeyer[9]等人提出了一种新颖的方法，他们将物体表面表达为 3D 占据场，并运用数值技术来确定射线与表面的准确交点。之后，通过隐式微分技术，他们精确地计算出这些交点的导数。这些交点位置信息随后被输入到一个神经 3D 纹理场中，该网络负责预测对应点的漫反射颜色。与此不同，Sitzmann 等人 [12]探索了一种更为间接的

神经隐式形状表示法。他们设计的系统能够为每一个连续的三维坐标位置输出特征向量和 RGB 颜色。而该技术的关键在于一个基于递归神经网络构建的可微分渲染函数，它通过沿每条射线进行迭代，从而精确地锁定物体的表面位置。

2.2. 新视图合成

在三维场景建模与渲染的进程中，新视图合成占据着举足轻重的地位。尤其是基于图像的视图合成技术，已成为计算机图形学和计算机视觉两大领域共同瞩目的核心议题。其核心思想是捕捉多个视角下的图像作为数据源，深入剖析这些图像中所包含的三维场景的几何构造、材料属性以及光照状况等因素，以便精确合成新视角下的视图。这一过程实际上是从一系列稀疏的二维图像中，构建并呈现出三维场景的新颖视图。早期的工作，通过从不同视角拍摄的图像中提取特征点并估计相机运动来重建场景的三维几何结构，]，这主要是在稀疏表示下重构场景。在此基础上，束调整 [14]和基于光照的方法 [4]考虑光和磁阻特性来合成逼真的图像，束调整方法通过优化相机参数和特征点的几何约束来改善场景重建的准确性。而基于光照的方法则考虑了光照条件对图像合成的影响，通过模拟光照和表面材质来提高合成图像的真实感。[7]方法通过融合多个局部光场以及它们对应的深度估计，生成高质量的全局光场表示，从而改善场景重建的准确性。Wiles[5]提出的 SynSin 端对端模型，学习特征的高分辨率点云来表示 3D 场景结构，使用一对卷积网络从输入图像中预测，可以在复杂的现实场景中从单幅图像合成视图。最近，神经渲染技术被引入到场景表示任务中，这启发了一系列研究来将 3D 场景建模为连续表示。场景表示网络 [13]首先将场景建模为 3D 坐标的函数，然后用于预测物体表面的交点和相应的发射颜色。跟随 SRN，神经辐射场（NeRF）[8]考虑场景中的体密度和视角相关的辐射颜色，并用简单但有效的多层感知器进行建模。场景中每个位置的输出与神经渲染技术相结合，以合成新视图。

2.3. 不确定性估计

在计算机视觉领域，不确定性估计的价值日益凸显，在多个研究方向中占据重要地位。对于神经网络而言，测量其不确定性不仅有助于提升模型输出的可解释性，还能显著降低因模型误判而引发的严重故障风险。有几种方法可以利用贝叶斯法则，将不确定性以模型参数或输出的概率分布形式来表达。其中贝叶斯神经网络（BNN）[10, 3]采用后验分布来呈现不确定性，而这常常需要借助诸如变分推理之类的近似手段。而 Dropout 变分推理 [2]则是通过对相同的输入执行多次推理，以此来评估带有 Dropout 层的模型所具有的不确定性。

在新视图合成领域，不确定性估计同样展现出了巨大的潜力。早期的研究工作已经对不确定性在新视图合成中的应用进行了初步探索，并取得了一定的成果。其中，NeRF-W[6]通过引入了不确定性来建模场景中的瞬态物体，从而实现了更为真实的视图合成效果。但其不确定性估计更多地关注于图像之间的差异，而对训练数据内部的噪声考虑不足。与此同时，另一项研究 S-NeRF[11]也采用了不确定性建模的方法。该研究利用变分推理对场景中的不确定性进行探索。虽然不确定性与预测误差之间呈现出了一定的相关性，但 S-NeRF 在图像质量上的表现并未能达到预期的效果，其合成的视图在边缘部分往往显得较为模糊，而且为了获取不确定性图，S-NeRF 还需要进行多次推理操作，相对应增加了计算的复杂性和时间成本。

3. 融合不确定性估计和 NeRF 的隐式三维建模方法设计与实现

本章首先详细阐述了 NeRF 模型的架构和渲染机制，接着探讨了如何将不确定性估计整合到模型中，以提高对场景的预测质量和鲁棒性。此外，还讨论了主动学习策略如何用于选择最具信息量的样本，优化模型性能。最后，展示了可视化界面的设计，使用户能够直观地与模型交互，简化三维建模过程。

3.1. 神经辐射场

NeRF 使用了体渲染中常用的吸收和发射模型，即场景中的每个点都设定为一个光源，不但可以吸收光线，本身也可以发射光线，将场景的几何、材质和照明等信息全部包含在内。即将场景建模为连续函数 $f(\theta)$，该函数输出发射辐射值和体密度。

具体而言，NeRF 将连续的场景表示为一个 5D 向量值函数，如图 3.2 所示，其输入是 3D 位置 $\mathrm{x} = (x, y, z)$ 和 2D 视角方向 (θ, φ)，输出的是体积密度 σ 和发射的颜色 $\mathrm{c} = (r, g, b)$。实际上，将 2D 视角方向表示为一个 3D 向量 $\mathrm{d} = d_x + d_y + d_z$。为了近似这个连续的 5D 场景表示，NeRF 使用一个 8 层 MLP 网络 $F_\Theta : (\mathrm{x}, \mathrm{d}) \rightarrow (\mathrm{c}, \sigma)$。

MLP 网络 F_Θ 首先对输入的 3D 坐标 x 进行处理，随后数据会流经 8 个全连接层。这些层中，每一层使用 ReLU 激活函数，并包含 256 个通道，最终会生成体积密度 σ 以及一个 256 维的特征向量。然后，此特征向量会与相机射线的视角方向 d 相融合。融合后的数据会通过一个额外的全连接层，该层采用 ReLU 激活函数并含有 128 个通道，最终输出与视角相关的 RGB 颜色值。

3.1.1 渲染阶段

在渲染阶段，神经辐射场（NeRF）使用了经典体积渲染方法，通过光线和场景点的采样预测颜色和密度，并将其累积到最终的合成图像中。在具体的实现步骤中，通过对采样点或途径点的颜色值进行累加来计算像素的最终颜色。这些采样点或途径点是指相机光线从相机中心 $\mathrm{o} \in \mathbb{R}^3$ 经过要计算颜色的像素坐标并穿过场景中的各个点。假设用 $\mathrm{r}(t) = \mathrm{o} + t\mathrm{d}$ 表示通过图像平面上给定像素的相机光线，其近界和远界分别为 t_n 和 t_f，则该像素的期望颜色 $C(\mathrm{r})$ 可以表示为：

$$C\left(r\right) = \int_{t_n}^{t_f} T\ \left(t\right) \sigma\left(r\left(t\right)\right) c\left(r\left(t\right),\ d\right) dt \quad (1)$$

$$T\ \left(t\right) = exp\ \left(- \int_{t_n}^{t} \sigma\left(\ r\left(\ s\ \right)\ \right) ds\ \right) \quad (2)$$

其中 $T(t)$ 表示光线从 t_n 到 t 的透明度，$\sigma(\mathrm{r}(t))$ 是体积密度，$\mathrm{c}(\mathrm{r}(t), \mathrm{d})$ 是在位置 $\mathrm{r}(t)$ 沿着方向 d 的发射颜色。

为了数值估计这个连续积分，NeRF 使用了求积法。尽管在确定性的求积方法常被用于渲染离散化的体素网格，但它仅在固定的离散位置查询 MLP，从而对分辨率的表示产生了限制。为了应对这一难题，NeRF 采纳了分层采样的策略。为了克服这个问题，NeRF 采用了分层采样方法。分层采样方法将区间 $[t_n, t_f]$ 等间隔地划分为 N 个子区间，并在每个子区间内均匀随机地采样一个样本，如下式所述：

$$t_i \sim u\left[t_n + \frac{i-1}{N}\left(t_f - t_n\right), t_n + \frac{i}{N}\left(t_f - t_n\right)\right] \quad (3)$$

由此，将上述积分表示为采样点的线性组合：

$$t_i \sim u\left[t_n + \frac{i-1}{N}\left(t_f - t_n\right), t_n + \frac{i}{N}\left(t_f - t_n\right)\right] \quad (4)$$

$$\begin{aligned} \hat{C}(r) &= \sum_{i=1}^{N_s} \alpha_i c\left(r(t_i)\right), \\ \alpha_i &= \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)\left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \end{aligned} \quad (5)$$

其中，$\delta_i = t_{i+1} - t_i$ 是相邻样本之间的距离，N_s 表示样本数。

3.1.2 位置编码

传统的 MLP 网络在学习和表示颜色、纹理和光照等高频数据细节方面存在困难。因为通常场景中的颜色纹理信息具有高频成分，直接使用 MLP 网络学习可能导致纹理表面变得模糊。为了解决这个问题，NeRF 引入了位置编码的概念。位置编码函数采用了类似正余弦周期函数的形式，以不同频域对位置进行编码，使得 MLP 网络能够同时学习场景中的高低频信息，提高对细节的捕捉能力。位置编码的作用类似于傅里叶变换，将低维空间的数据输入映射到高维空间，增加网络对高频数据的敏感性。神经辐射场利用位置编码将输入的空间位置 $x = (x, y, z)$ 和观察方向 (θ, φ) 映射到高维空间后再传递给网络，

能够有效地拟合包含高频变化的数据，提高生成图像的清晰度。编码函数如下：

$$\gamma(p) = \left(\sin(2^0\pi \mathrm{p}), \cdots \ \cos(2^{L-1}\pi \mathrm{p})\right) \tag{6}$$

这个函数 $\gamma(\cdot)$ 分别应用于输入 x 中的三个坐标值（这些坐标值已经被归一化到 $[-1,1]$ 区间内），以及笛卡尔视角方向单位向量 d 的三个分量（这些分量通过构造限制在 $[-1,1]$ 区间内）。通过实验，设置 $L=10$ 用于 $\gamma(x)$ 的维度，以及 $L=4$ 用于 $\gamma(d)$ 的维度时效果最好。

3.1.3 分层体素采样

由于一条射线上大部分区域都是空区域或被遮挡的区域，对最终颜色的贡献值较小，按照公式 3-2 的均匀采样进行渲染效果较差，所以 NeRF 采用分层体素渲染方式，通过对不同区域分别进行粗采样和细采样的方式来减少计算开销。

在粗采样阶段，神经辐射场使用较为稀疏的采样点，在起点和终点之间均匀采样第一组 N_c 个采样点。这些粗采样的采样点用于计算体素的密度和颜色值。对于得到的粗采样点，将公式（5）合成颜色 $\hat{C}(r)$ 重写为沿射线所有采样颜色 c_i 的加权和的形式：

$$\hat{C}(r) = \sum_{i=1}^{N_c} w_i c_i, \quad w_i = T_i(1-\exp(-\sigma_i\delta_i)) \tag{7}$$

$$\hat{w}_j = \frac{w_i}{\sum_{j=1}^{N_c} w_j} \tag{8}$$

将权重按照公式 3-6 将这些权重标准化后，会在射线上产生分段常数概率密度函数（PDF）。使用逆变换采样从这个分布中采样第二组 N_f 个采样点，然后在第一组和第二组采样点的并集上对细网络进行评估，并使用方程 3-3 计算射线的最终渲染颜色 $\hat{C}_f(r)$。

在此基础上，NeRF 通过最小化真实 RGB 图像 $\{I_i\}_{i=1}^N$ 与渲染像素颜色之间的平方重建误差来优化连续函数 F_θ。

$$\sum_i \left\|\hat{C}_c(r_i) - C(r_i)\right\|_2^2 + \left\|\hat{C}_f(r_i) - C(r_i)\right\|_2^2 \tag{9}$$

其中，r_i 表示采样的光线，$C(r_i)$、$\hat{C}_c(r_i)$、$\hat{C}_f(r_i)$ 分别对应于真实值、粗模型预测和细模型预测。

3.2. 不确定性估计的融合

当前研究表明，NeRF 在处理单个或少量输入视图时，其泛化能力颇显不足。当面对的场景未能完全呈现时，传统的 NeRF 框架会倾向于预测未观测区域的体积密度为零，导致出现平凡解。这往往导致结果趋于平庸，缺乏深度与真实感。因此，本文实现将场景中每个位置的发射辐射值建模为高斯分布，而不是单个值。预测的方差可以反映与特定位置相关的随机不确定性。通过这种方式，模型被迫在未观察到的区域提供更大的方差，而不是陷入平凡解。

3.2.1 模型架构

引入不确定性估计后，模型如图2所示，其输入仍是经过位置编码的 3D 位置 $x=(x,y,z)$ 和 2D 视角方向 (θ,φ)，输出为辐射颜色的均值和方差。具体而言，我们定义相机光线上某点 $r(t)$ 的辐射颜色遵循由均值 $\bar{c}(r(t))$ 和方差 $\bar{\beta}^2(r(t))$ 参数化的高斯分布。遵循贝叶斯神经网络的先前研究，我们将模型输出作为均值，并在 NeRF 的 MLP 网络中添加额外的分支来建模方差，如下所示：

$$[\sigma, f, \beta^2(r(t))] = \mathrm{MLP}_{\theta_1,\theta_3}(\gamma_x(x)) \tag{10}$$

$$\hat{C} = \mathrm{MLP}_{\theta_2}(f, \gamma_d(d)) \tag{11}$$

Softplus 函数具有非负性、平滑性、单调递增和数值稳定性，其定义如公式所示，因此它能够通过提供一个非零的最小方差并允许方差随着输入的增加而单调递增，有助于模型合理地表达场景的不确定性。本文采用 Softplus 函数来生成一个有效的方差值，确保模型输出的方差非负且平滑：

$$\mathrm{Softplus}(x) = \log(1+\exp(x)) \tag{12}$$

$$\bar{\beta}^2(r(t)) = \beta_0^2 + \log(1+\exp(\beta^2(r(t)))) \tag{13}$$

其中，β_0^2 是一个常数项，确保所有位置都具有最小的方差。

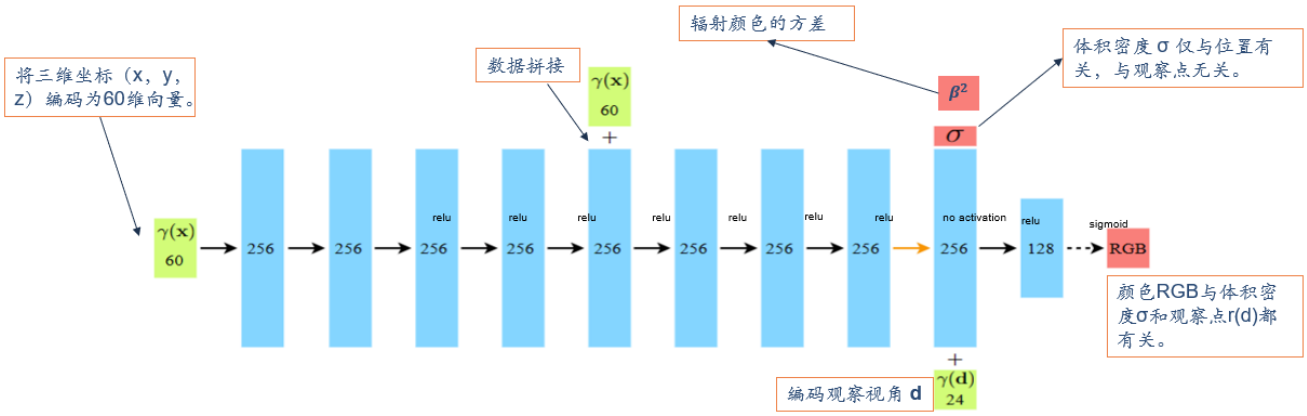

图 2. 模型架构图

3.2.2 渲染过程

在渲染过程中，引入不确定性后的新神经辐射场仍然可以通过体渲染技术高效实现。NeRF 框架的设计理念提供了两个重要的基础：首先，特定位置的辐射颜色仅受其三维坐标的影响，这确保了不同位置的辐射特性分布相互独立；其次，体积渲染可以通过沿光线路径的采样点进行线性组合来近似实现。基于这些前提，将位置 $r(t)$ 处的辐射颜色表示为一个高斯分布 $c(r(t)) \sim N(\bar{c}(r(t)), (\bar{\beta}^2)(r(t)))$，所以沿该光线路径的渲染值将自然地也遵循高斯分布：

$$\hat{C}(r) \sim N\left(\sum_{i=1}^{N_s} \alpha_i \bar{c}(r(t_i)), \sum_{i=1}^{N_s} \alpha_i (\bar{\beta}^2)(r(t_i))\right) \sim N(\bar{C}(r), (\bar{\beta}^2)(r)), \tag{14}$$

3.2.3 优化与损失

在 3D 场景中，两条不同的光线相交于同一点的概率是很低的。即使在复杂的场景中，这种交集也不太可能发生。假设在任何一个给定的训练批次中，场景中的任何一个特定位置最多只会被采样一次。因此，每个采样点都是独立的，所以从这些点渲染出来的光线对应的颜色分布也被认为是相互独立的。

通过这种方式，我们可以通过最小化来自批次 B 的光线 $r_i{}_{i=1}^{N}$ 的负对数似然来优化模型：

$$\begin{aligned}\min_{\theta} -\log p_\theta(B) &= -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(C(r_i)) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|C(r_i)-\bar{C}(r_i)\|_2^2}{2\bar{\beta}^2(r_i)} + \frac{\log \bar{\beta}^2(r_i)}{2}\right).\end{aligned} \tag{15}$$

然而，直接最小化公式 (3-13) 的目标函数可能导致次优结果，其中不同采样点在光线中的权重 α_i 趋向于相互靠拢。为了避免这一问题，我们在损失函数中引入了一个额外的正则化项，目的是促使体密度分布更加稀疏化。通过这种方式，我们能够有效地控制场景中非零密度部分的占比，从而提高物体表面清晰度，优化新视角合成的图像质量。具体如下所示：

$$\begin{aligned}L^{\text{uct}} = \frac{1}{N}\sum_{i=1}^{N}\Bigg(&\frac{\|C(r_i)-\bar{C}(r_i)\|_2^2}{2\bar{\beta}^2(r_i)} + \frac{\log \bar{\beta}^2(r_i)}{2} \\ &+ \frac{\lambda}{N_s}\sum_{j=1}^{N_s}\sigma_i(r_i(t_j))\Bigg).\end{aligned} \tag{16}$$

遵循原始的 NeRF 框架并优化两个并行网络。为了简化优化的难度，我们只在精细模型中采用不确定性分支，并保持粗糙模型与原始模型相同。最终的损失函数如下：

$$L = L^{\text{uct}}(C(r), \bar{C}^f(r)) + \frac{1}{N}\sum_{i=1}^{N} \|\hat{C}_c(r_i) - C(r_i)\|_2^2. \quad (17)$$

综上所述，本文实现模型的训练过程如图 3.4 所示，它首先接收输入数据，包括空间坐标和方向向量，然后经过位置编码，通过网络得到光线上采样点的颜色、体密度和方差。模型采用两种多层感知器网络，一个“细致”网络用于产生精细的预测值 $C^r(r)$，另一个“粗致”网络用于产生粗略的预测值。这两种预测值与真实值 $C(r)$ 结合，通过损失函数进行优化，从而调整网络参数。

4. 实验验证与效果展示

本章将详细阐述了实验过程中所使用的数据集、数据预处理步骤、评价指标、对比方法，并且展示 3D 渲染效果，对对比实验结果进行了深入分析。通过多方面的比较和剖析，充分验证了本文所实现的融合不确定性估计与 NeRF 的隐式三维建模方法的卓越性能与优势。

4.1. 实验设置

在本文的实验部分，本文深入探索了模型性能与不同数量输入数据之间的关系。为了更全面、客观地评估模型的表现，本文特意将本文实现的模型与其他主流模型，尤其是 NeRF 模型，进行了详尽的对比分析。在评估过程中，本文不仅依赖了单一的图像质量指标，而是综合运用了 PSNR（Peak Signal-to-Noise Ratio，峰值信噪比）、SSIM（Structural Similarity Index，结构相似性指数）和 LPIPS（Learned Perceptual Image Patch Similarity，学习的感知图像块相似性）等多项指标，力求从多个维度全面的反映模型的性能。为了更贴近实际应用场景，本文特意采用了不同比例的训练样本，对合成场景和真实场景都进行了深入的定性分析。

在深入实验细节阐述之前，首先概述实验环境配置。在硬件配置上，由于作者电脑的显卡配置较低显存不足，所以本文选用了实验室的高性能服务器。这台服务器搭载了 Intel® 13th Gen Intel(R) Core(TM) i9-13900K CPU，以及具备 24GB 显存的 Nvidia GeForce GTX 3090 图形显卡，操作系统为稳定且功能强大的 Ubuntu 22.04.3 LTS Desktop，确保了实验的高效运行。在软件环境上，本文使用了 Visual Studio Code 作为开发工具，其可以方便与服务器在同一局域网下连接，同时选用了 CUDA 11.4 作为计算库，以提升大规模并行计算的效率。深度学习框架方面，我选择了功能强大、灵活性高的 Pytorch 1.11.0，而编程语言，则是选择了广泛应用的 Python3.8。值得一提的是，本研究中所有的神经网络结构、优化算法以及数据加载策略，都是基于 PyTorch 框架实现的，并且所有的实验都是在 GPU 加速环境下完成的。

在接下来的内容中，将详细阐述本次研究所采用的数据集及其数据预处理操作、介绍三种性能指标、与其他模型的对比方法，并会根据实验结果展示主动学习策略的优异性。

4.2. 实验数据

4.2.1 **数据集**

本文使用的数据来自现有 NeRF 数据集、LLFF 数据集和自定义数据集。

NeRF 数据集是合成数据集，通过计算机图形学技术生成的虚拟场景数据。其包含 8 个具有复杂几何形状和逼真非兰伯特材料的合成对象。每个场景有 100 个视角用于训练和 200 个视角用于测试，所有图像均为 800×800 的分辨率。合成数据集的优点在于可以精确控制场景和生成真实深度信息，这有助于模型学习准确的辐射传输和场景重建。但其可能无法完全捕捉真实世界中的复杂性和变化。

LLFF 是真实数据集，一个由手机拍摄的 8 个复杂场景的真实数据集。每个场景包含 20–60 个分辨率为 1008×756 的图像，其中 $\frac{1}{8}$ 的图像保留用于测试。其优点在于能够提供更真实的场景和变化，并且模型在真实数据集上的性能更具可靠性。但是真实数据集的采集和标注过程可能更加复杂和耗时，并且可能难以精确控制光照和深度信息。

自定义数据集是由作者拍摄的两个场景构成的

真实数据集，每个场景包含 50–60 个分辨率为 1280×720 的图像。

4.2.2 数据预处理

NeRF 数据集和 LLFF 数据集从 NeRF 官方下载获得，其中已经做好了数据采集、特征匹配与位姿估计工作，下面主要介绍一下自定义数据集的数据预处理工作。

（1）数据采集

自定义数据集通过使用手机在场景中不同位置和角度下的拍摄获得，本实验对两个场景进行数据采集，分别为对我个人和对我的水杯进行数据采集工作。

采集过程中特别要确定拍摄的视角和观察方向。视角的选择应该尽可能地覆盖整个场景，包括不同位置和角度的图像，以获得全面的场景信息。可以通过移动手机、调整手机的高度和角度来改变视角。

（2）特征匹配与位姿估计

本文采用 COLMAP（Structure-from-Motion and Multi-View Stereo）这一工具包对采集到的自定义数据集进行特征匹配与位姿估计，它提供了一个完整的流程，用于从图像集合中恢复出场景的三维结构和相机的位姿信息。

首先，COLMAP 会从每个图像中提取特征点。它支持多种特征提取算法，如 SIFT、SURF、ORB 等，这些算法能够检测具有良好不变性和可重复性的特征点。COLMAP 会计算每个特征点的坐标和尺度，并生成相应的特征描述子。

接下来，COLMAP 会利用特征描述子进行特征匹配。它会对不同图像之间的特征点进行匹配，建立它们之间的对应关系。COLMAP 使用一些经典的匹配算法，比如最近邻匹配、基于距离阈值的匹配等。通过比较特征描述子的相似性，COLMAP 能够找到最佳的匹配对。

为了提高匹配的准确性，COLMAP 会对匹配结果进行筛选和过滤。例如，它可以使用距离比例测试来排除不可靠的匹配。该测试会比较每个特征点的两个最佳匹配之间的距离比，如果比值超过一个阈值，则将该匹配筛除。通过特征匹配，COLMAP 可以获取图像之间的相对姿态信息。

利用这些匹配信息，COLMAP 会进行初始的相机位姿估计。它使用了基于几何一致性的 RANSAC 算法，通过随机采样和模型拟合的方式，找到最佳的匹配子集，并计算相机的初始位姿。

相机位姿估计后，就可以从位姿估计结果中提取相机的内外参数，包括相机的焦距、图像的宽度和高度、相机的位置和朝向等信息。并对场景进行稀疏重建。

（3）转换成 LLFF 数据集格式

为了使自定义数据集满足本文模型提出模型的需要，要将 Colmap 预测的相机位姿转换为 LLFF 数据集，首先从位姿估计结果中提取相机的内外参数。然后准备需要转换的图像，并确保它们与相机位姿对应。接下来，使用相机参数计算每个图像样本点的观察方向向量，并根据相机位置、朝向和图像中心点的像素坐标计算样本点在相机坐标系中的像素坐标。最后，将图像、相机参数、观察方向向量和像素坐标等信息组织成 LLFF 数据集的格式。

（4）视角计算

对于 LLFF 数据集，使用相机参数和像素坐标信息，通过逆投影将像素坐标转换为相机坐标系中的射线方向。然后，将射线方向归一化为单位向量，得到观察方向向量。计算得到的观察方向向量表示从相机位置观察样本点的视角。

4.3. 实验结果分析

4.3.1 不同比例样本下的对比

表1展示了所实现方法与其他三维隐式建模方法在使用不同比例的训练样本对合成场景和真实场景进行定性分析的结果。

以合成场景为例，当使用全部图片进行训练时，NeRE 方法在 PSNR 上达到了 31.01，SSIM 为 0.947，LPIPS 为 0.081，显示出优秀的性能。然而，当训练样本数量减少到 10 张时，NeRE 的 PSNR 下降到 28.04，SSIM 为 0.866，LPIPS 为 0.134，性能有所下降。进一步减少到 5 张图片时，性能下降更为明显，PSNR 降至 21.14，SSIM 为 0.835，LPIPS 增至 0.192。这种趋势在真实场景中也得到了类似的体

现。

相比之下，本文实现的方法（Ours）在融合了不确定性估计后，即使在训练样本数量较少的情况下，也展现出了较为稳定和出色的性能。在合成场景中，使用 5 张图片训练时，Ours 的 PSNR 达到了 23.23，SSIM 为 0.866，LPIPS 为 0.185，这表明其在样本效率和鲁棒性方面具有显著优势。这在资源受限的实际应用场景中尤为重要，尤其是在需要快速部署和适应新环境的情境下。

此外，本文实现的方法（Ours）在真实场景中的表现也同样令人印象深刻。在全部图片训练的情况下，Ours 的 PSNR 为 25.96，SSIM 为 0.835，LPIPS 为 0.213，与合成场景相比，虽然面临更加复杂的现实世界条件，但性能依然保持在较高水平。这一点在减少训练样本数量时尤为突出，显示了本文方法在处理真实世界数据时的强大适应性和鲁棒性。

4.3.2 主动学习性能验证

5. 主动学习策略的实验验证

在进行主动学习策略的研究中，本实验旨在通过与两种启发式方法的对比，验证所提出的策略的有效性。实验采用了一种近似的方法，即在训练集中保留了大部分图像，并将它们作为候选样本。具体来说，基线方法包括了两种不同的候选图像捕获策略：NeRF + Random（随机捕获候选图像）和 NeRF + FVS（最远视角采样，即选择与当前训练集摄像机位置距离最远的候选图像）。

实验设计了两种不同的设置来评估不同策略的性能。在设置 I 中，我们从 4 个初始观察开始，并在后续的 40K、80K、120K 和 160K 迭代中分别获得了 4 个额外的观察。而在设置 II 中，我们从 2 个初始观察开始，并在相同的迭代次数中分别获得 2 个额外的观察。这样的设计旨在模拟不同的数据采集预算和迭代次数对学习效果的影响。

表2和 图3展示了在连续学习（Continuous Learning, CL）方案下的结果，这一方案假设了充足的时间和计算资源。从结果中可以明显看出，与启发式方法相比，本文实现的主动学习策略能够更有效地捕获最具信息量的输入样本，这些样本对于从未观察到的区域合成视角的贡献更为显著。此外，尽管本文实现的模型需要额外的训练时间（2.2 小时相对于 2 小时），但这种额外的训练成本相对较小，且能够带来显著的性能提升。

为了进一步验证模型的性能，我们还采用了结合贝叶斯估计（Bayesian Estimation, BE）的方法进行了测试。如表2所示，采用贝叶斯估计的模型在时间消耗上可以节省高达 75%，这表明了贝叶斯估计在提高效率方面的显著优势。尽管在连续学习方案下，贝叶斯估计的性能可能略逊一筹，但它仍然能够合成出合理的图像，并且在某些情况下，其性能甚至可以与启发式方法相媲美。

6. 总结与展望

6.1. 本文主要工作和论文总结

在本文中，我们实现了融合不确定性估计和 NeRF 的隐式三维建模方法。该方法提高了三维场景重建的准确性和可靠性，特别是在数据稀缺的情况下。具体来说，通过引入高斯分布对场景辐射颜色进行建模，有效提升了模型在数据稀缺情况下的泛化能力和预测质量。同时，结合主动学习策略，优化了模型性能，通过选择最具信息量的新视角图像，最小化额外资源消耗提升渲染质量。实验结果表明，该方法在合成和真实场景中均表现出色，尤其适用于快速部署和适应新环境的应用场景。此外，本文开发的 PyQT 可视化界面极大地简化了模型的使用和推广，为非专业用户提供了便捷的操作体验。

6.2. 未来工作展望

本文实现的算法相较于其他方法在 3D 建模性能上有所优势，在准确性和效率上均有提升，并且提供了场景的可解释性，但仍存在一些局限性。

第一，渲染速度较慢。由于使用多层感知机（MLP）对场景进行建模，每次渲染都需要进行大量的前向传播计算，这导致渲染速度较慢。在某些情况下，渲染一幅图像可能需要数小时甚至更长时间。这种长时间的渲染过程不利于需要实时或近实时渲染的应用场景。第二，对训练数据的依赖较大。尽管

表 1. 不同样本场景下的模型表现表

方法	(a) 合成场景			(b) 真实场景		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
数据集 1，使用全部图片训练						
SRN	22.26	0.846	0.170	22.84	0.668	0.378
LLFF	24.88	0.911	0.114	24.13	0.798	0.212
NeRF	31.01	0.947	0.081	26.50	0.811	0.250
Ours	30.45	0.954	0.072	25.96	0.835	0.213
数据集 2，使用 10 张图片训练						
NeRF	28.04	0.866	0.134	23.36	0.791	0.280
DietNeRF	28.42	0.891	0.087	-	-	-
Ours	28.51	0.932	0.090	23.96	0.803	0.260
数据集 3，使用 5 张图片训练						
NeRF	21.14	0.835	0.192	21.67	0.689	0.350
Ours	23.23	0.866	0.185	22.03	0.712	0.292

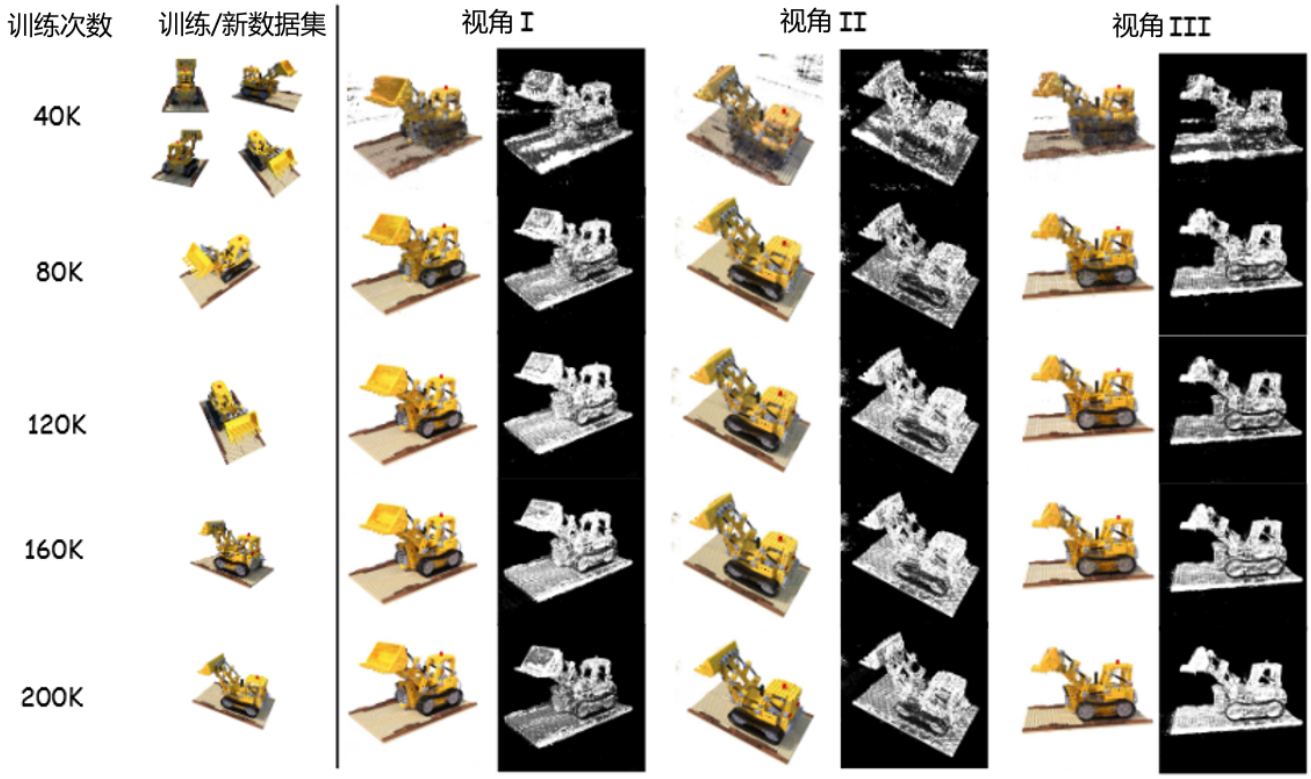

图 3. 主动迭代的定性结果图

本文实现的方法在少量训练样本下表现出较好的性能，但数据集的采集和预处理过程耗时耗力，且需要特定的技术知识。第三，对动态场景的建模不足。本文模型主要针对静态场景的三维建模，对于动态场景（如包含移动物体或时间变化的场景）的重建能力有限。本文方法在处理动态场景时可能存在不

表 2. 主动学习策略设置下的定量结果表

方法	时间	(a) 合成场景			(b) 真实场景		
		PSNR	SSIM	LPIPSJ	PSNR	SSIM	LPIPSJ
设置 1, 共 20 个观测值:							
NeRF+Rand	2.0h	24.25	0.734	0.207	20.65	0.532	0.312
NeRF+FVS	2.0h	26.00	0.812	0.144	22.41	0.710	0.299
Ours-BE	30min	25.67	0.778	0.169	21.86	0.644	0.303
Ours-CL	2.2h	26.24	0.856	0.124	23.12	0.765	0.292
NeRF†	2.0h	28.04	0.910	0.134	23.36	0.791	0.280
设置 2, 共 10 个观测值:							
NeRF+Rand	1.0h	18.36	0.642	0.251	18.49	0.478	0.355
NeRF+FVS	1.0h	19.24	0.735	0.227	20.02	0.633	0.344
Ours-BE	16min	18.25	0.611	0.256	18.67	0.451	0.367
Ours-CL	1.1h	20.01	0.832	0.204	20.14	0.664	0.325
NeRF†	1.0h	21.14	0.835	0.192	21.67	0.689	0.350

足，难以有效捕捉和重建动态信息。

7. 致谢

本科生涯即将画上句号，我静静回想起大学四年内的每一个日夜。从懵懂无知的新生到即将开启研究生生活的大四毕业生，这一路走来，我经历了无数的挑战与成长。借此机会向身边关心我的每个人表达感谢。

首先，衷心感谢我的父母和家人。你们一直以来的无私支持和鼓励是我不断前行的动力。无论是生活中的点滴关怀，还是在我遇到困难时给予的坚定支持，你们都让我感受到无尽的温暖和力量！

其次，我要特别感谢我的指导老师陈蕾老师。在毕业设计期间，陈老师一直为我提供精心的引导和协助。从选题、确定研究方法，到论文的撰写与修订，陈老师始终为我提供了巨大的支持和鼓舞。陈老师深厚的学识、严密的学术态度和科研工作的热忱，都深深打动了我，让我收获颇丰。同时，我也要感谢助教韩松成学长，在我遇到技术难题和困惑时，韩学长总是耐心解答，提供宝贵的建议和指导，帮助我顺利克服了一个又一个难关。

最后，我要感谢我的导员韩烁老师和一直在我身旁的朋友们。在这段时间里，你们的陪伴和支持让我在繁忙的学习和科研之余，感受到友谊的温馨和快乐。无论是一起讨论学术问题，还是在生活中互相帮助，你们的友情让我倍感珍惜。

希望未来的生活里，我能继续努力，不负青春年华！

参考文献

[1] A. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. arXiv: Graphics,arXiv: Graphics, Dec 2015. 2

[2] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Cornell University - arXiv,Cornell University - arXiv, Jun 2015. 3

[3] I. Kononenko. Bayesian neural networks. Biological Cybernetics, 61(5):361–370, Sep 1989. 3

[4] M. Levoy and P. Hanrahan. Light field rendering. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH ' 96, Jan 1996. 3

[5] C. Liu, W. Freeman, R. Szeliski, and S. B. Kang. Noise estimation from a single image. In 2006 IEEE Computer Society Conference on Computer Vision and

Pattern Recognition - Volume 1 (CVPR'06), Jul 2006. 3

[6] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2021. 3

[7] B. Mildenhall, P. Srinivasan, R. Ortiz-Cayon, N. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, and L. Light. Local light field fusion. 3

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, page 405–421. Jan 2020. 3

[9] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Supplementary material for differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. 2

[10] D. Nosek and J. Noskov
' a. On bayesian analysis of on-off measurements. 3

[11] J. Shen, A. Ruiz, A. Agudo, and F. Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In 2021 International Conference on 3D Vision (3DV), Dec 2021. 3

[12] V. Sitzmann, M. Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. Neural Information Processing Systems,Neural Information Processing Systems, Jun 2019. 2

[13] V. Sitzmann, M. Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. Neural Information Processing Systems,Neural Information Processing Systems, Jun 2019. 3

[14] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle Adjustment —A Modern Synthesis, page 298–372. Jan 2000. 3