# Neural Sparse Voxel Fields

论文地址：<https://proceedings.neurips.cc/paper/2020/file/b4b758962f17808746e9bb832a6fa4b8-Paper.pdf>

Github：<https://github.com/facebookresearch/NSVF>

## 摘要

使用经典计算机图形技术**对真实世界场景进行逼真的自由视角渲染**具有挑战性，因为它需要**捕获详细外观和几何模型的困难步骤**。最近的研究通过学习在没有 3D 监督的情况下隐式编码几何和外观的场景表示，证明了有希望的结果。然而，现有方法在实践中经常表现出由于网络容量有限或难以找到相机光线与场景几何体的准确交点而导致的**模糊渲染**。从这些表示中合成高分辨率图像通常需要**耗时的光线行进**。在这项工作中，我们引入了神经稀疏体素场 (NSVF)，这是一种用于快速和高质量自由视角渲染的新神经场景表示。 NSVF 定义了一组以**稀疏体素八叉树组织的体素有界隐式场**，以对每个单元格中的局部属性进行建模。我们仅从一组构成的 RGB 图像中，通过可微分的光线行进操作逐步学习潜在的体素结构。使用稀疏体素八叉树结构，可以通过跳过不包含相关场景内容的体素来加速渲染新视图。我们的方法在推理时通常比最先进的方法（即 NeRF（Mildenhall 等人，2020 年））快 10 倍以上，同时获得更高质量的结果。此外，通过利用显式稀疏体素表示，我们的方法可以轻松应用于场景编辑和场景合成。我们还展示了几个具有挑战性的任务，包括多场景学习、移动人的自由视角渲染和大规模场景渲染。

## 介绍

计算机图形学中的逼真渲染具有广泛的应用，包括混合现实、视觉效果、可视化，甚至计算机视觉和机器人导航中的训练数据生成。从任意角度逼真地渲染真实世界场景是一项巨大的挑战，因为通常无法像在高预算的视觉效果制作中那样获得高质量的场景几何图形和材料模型。因此，研究人员开发了**基于图像的渲染 (IBR) 方法，将基于视觉的场景几何建模与基于图像的视图插值相结合（**Shum 和 Kang，2000；Zhang 和 Chen，2004；Szeliski，2010）。尽管取得了重大进展，但 IBR 方法仍然具有**次优渲染质量和对结果的有限控制**，并且通常是特定于场景类型的。为了克服这些限制，最近的工作采用了**深度神经网络来隐式地学习场景表示，从带有或不带有粗几何的 2D 观察中封装几何和外观**。这种神经表示通常与 3D 几何模型相结合，例如体素网格（Yan 等人，2016 年；Sitzmann 等人，2019a；Lombardi 等人，2019 年）、纹理网格（Thies 等人，2019 年；Kim 等人al., 2018; Liu et al., 2019a, 2020)、多平面图像（Zhou et al., 2018; Flynn et al., 2019; Mildenhall et al., 2019）、点云（Meshry et al., 2019；Aliev 等人，2019 年）和隐函数（Sitzmann 等人，2019b；Mildenhall 等人，2020 年）。

与大多数显式几何表示不同，神经隐式函数是平滑、连续的，并且理论上可以实现高空间分辨率。 然而，现有方法在实践中经常表现出由于网络容量有限或难以找到相机光线与场景几何体的准确交点而导致的**模糊渲染**。从这些表示中合成高分辨率图像通常需要**耗时的光线行进**。 此外，使用这些神经表示**编辑或重新合成 3D 场景模型并不简单**。

在本文中，我们提出了神经稀疏体素场 (NSVF)，这是一种用于快速和高质量自由视点渲染的新隐式表示。 NSVF 不是使用单个隐函数对整个空间进行建模，而是由一组以稀疏体素八叉树组织的体素有界隐式字段组成。具体来说，我们在**体素的每个顶点分配一个体素嵌入，并通过在相应体素的八个顶点处聚合体素嵌入来获得体素内部查询点的表示**。这进一步通过多层感知器网络 (MLP) 来预测该查询点的几何形状和外观。我们的方法可以通过可微分的光线行进操作从场景的一组 2D 图像中逐步学习 NSVF，从粗到细。在训练期间，**不包含场景信息的稀疏体素将被修剪**，以允许网络专注于具有场景内容的体积区域的隐函数学习。使用稀疏体素，可以通过跳过没有场景内容的空体素来大大加快推理时的渲染速度。

我们的方法在推理时通常比最先进的方法（即 NeRF（Mildenhall 等人，2020 年））快 10 倍以上，同时获得更高质量的结果。 我们在各种具有挑战性的任务上广泛评估了我们的方法，包括多对象学习、动态和室内场景的自由视点渲染。 我们的方法可用于编辑和合成场景。 总而言之，我们的技术贡献是：

* 我们提出了由一组体素有界隐式场组成的 NSVF，其中对于每个体素，体素嵌入被学习以编码局部属性以实现高质量渲染；
* NSVF 利用稀疏体素结构实现高效渲染；
* 我们引入了一种渐进式训练策略，该策略以端到端的方式从一组有姿势的 2D 图像中通过可微分的光线行进操作有效地学习底层稀疏体素结构。

## 背景

现有的神经场景表示和神经渲染方法通常旨在学习将空间位置映射到隐式描述场景的局部几何形状和外观的特征表示的函数，其中可以使用计算机图形学中的渲染技术合成该场景的新视角。为此，渲染过程以可微分的方式制定，以便可以通过最小化渲染和场景的 2D 图像之间的差异来训练对场景表示进行编码的神经网络。 在本节中，我们将描述使用隐式场的现有表示和渲染方法及其局限性。

### Neural Rendering with Implicit Fields 隐式场的神经渲染

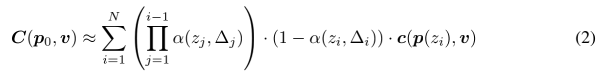
让我们将场景表示为隐函数 ，其中 θ 是底层神经网络的参数。 该函数描述了场景颜色 c 及其在空间位置 p 和光线方向 v 处的概率密度 ω。给定位置 处的针孔相机，我们通过从相机拍摄光线来把大小为 H × W 的二维图像渲染成 3D 场景。 因此，我们评估体积渲染积分以计算相机光线的颜色为：



请注意，为了鼓励场景表示是多视角一致的，ω 被限制为仅 p(z) 的函数，而 c 将 p(z) 和 v 作为输入来建模与视角相关的颜色。 评估该积分的不同渲染策略是可行的。

**Surface Rendering**。 基于表面的方法 (Sitzmann et al., 2019b; Liu et al., 2019b; Niemeyer et al., 2019) 假设 ω(p(z)) 是狄拉克函数 δ(p(z)) p(z∗ )) 其中 p(z∗) 是相机光线与场景几何的交点。

**Volume Rendering.**（Lombardi 等人，2019 年；Mildenhall 等人，2020 年）通过在每条相机光线上密集采样点并将采样点的颜色和密度累积到 2D 图像中来估计方程(1)中的积分 C(p0, v)。例如，最先进的方法 NeRF（Mildenhall 等人，2020 年）估计 C(p0, v) 为：



其中  ，然后 和则是采样点的颜色和体积密度。

### 现在方法的局限性

对于表面渲染，**至关重要的是找到一个准确的表面**，使学习到的颜色在多视角中保持一致，这对训练收敛很困难且有害，从而导致渲染中的模糊。**体积渲染方法需要沿光线采样大量点以进行颜色累积，以实现高质量渲染**。然而，像 NeRF 那样沿射线评估每个采样点是低效的。例如，NeRF 渲染 800 × 800 图像需要大约 30 秒。**我们的主要观点是，尽可能防止在没有相关场景内容的空白空间中采样点很重要**。尽管 NeRF 沿射线执行重要性采样，但由于为每条射线分配了固定的计算预算，它无法利用这个机会来提高渲染速度。我们受到经典计算机图形技术的启发，例如边界体积层次结构（BVH，Rubin 和 Whitted，1980）和稀疏体素八叉树（SVO，Laine 和 Karras，2010），它们旨在以稀疏层次结构对场景进行建模光线追踪加速。在这种编码中，空间位置的局部属性仅取决于空间位置所属的叶节点的局部邻域。在本文中，我们展示了如何在 3D 场景的神经网络编码隐式字段中使用分层稀疏体积表示，以实现详细编码和高效、高质量的可微体积渲染，即使是大型场景也是如此。

## Neural Sparse Voxel Fields

在本节中，我们将介绍神经稀疏体素场 (NSVF)，这是一种**将神经隐式场与显式稀疏体素结构相结合的混合场景表示**。 NSVF 不是将整个场景表示为单个隐式场，而是由一组以稀疏体素八叉树组织的体素有界隐式场组成。 在下文中，我们描述了 NSVF 的构建块 - 一个体素有界隐式场（第 3.1 节） - 随后是 NSVF 的渲染算法（第 3.2 节）和渐进式学习策略（第 3.3 节）。

### Voxel-bounded Implicit Fields 体素有界隐式场

我们假设场景的相关非空部分包含在一组稀疏（有界）体素 ，并且场景被建模为一组体素有界隐函数： 其中p ∈ Vi 。 每个  都被建模为具有共享参数 θ 的多层感知器 (MLP)：



这里 c 和 表示 3D 点 p 的颜色和密度，v 表示光线方向，在点 p的表示  被定义为：



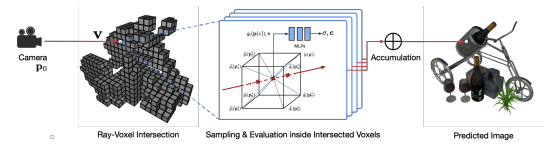
其中  表示 Vi 的八个顶点，是存储在每个顶点的特征向量。另外，χ(.)是指三线性插值，ζ(.)是后处理函数。 在我们的实验中，ζ(.) 是由 (Vaswani et al., 2017; Mildenhall et al., 2020) 提出的位置编码。

与之前大多数工作一样使用点 p 的 3D 坐标作为 的输入相比，在 NSVF 中，特征表示 (p) 由相应体素的八个体素嵌入聚合，其中区域特定信息（例如几何、 材料、颜色）可以嵌入。 它显着简化了后续 的学习，并促进了高质量的渲染。

**Special Cases.** NSVF 包含两类早期作品作为特例。 (1) 当 = 且ζ(.) 为位置编码时， ，这意味着 NeRF (Mildenhall et al., 2020) 是 NSVF 的特例。 (2) 当 ζ(.) 和 是恒等函数时，我们的模型等价于使用显式体素存储颜色和密度的模型，例如，Neural Volumes (Lombardi 等人，2019 年）。

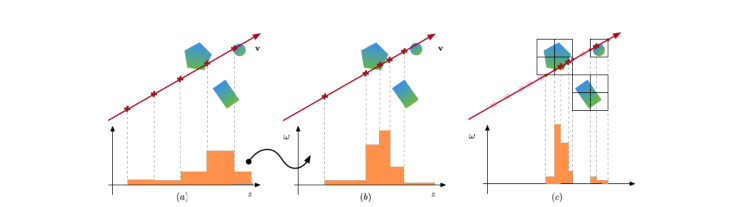
### 体渲染

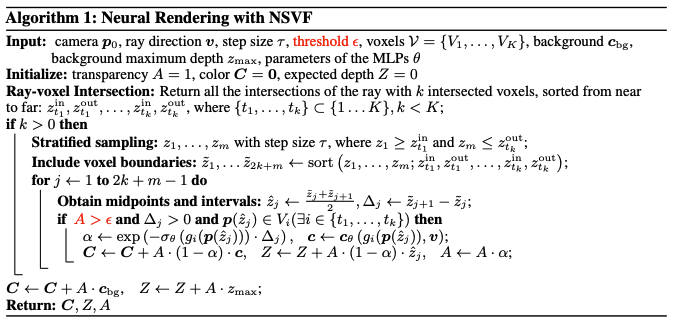
NSVF 在任何点 p ∈ V 对场景的颜色和密度进行编码。 与渲染模拟整个空间的神经隐式表示相比，渲染 NSVF 效率更高，因为它避免了空白空间中的采样点。 渲染分两步进行：（1）光线-体素相交； (2) 在体素内的光线行进。 我们在附录图 8 中说明了整个 pipeline，即下图所示



**Ray-voxel Intersection**. 我们首先对每条光线应用轴对齐边界框相交测试（AABB 测试）（Haines，1989）。 它通过比较从光线原点到体素的六个边界平面中的每一个的距离来检查光线是否与体素相交。 **AABB 测试非常有效，尤其是对于分层八叉树结构（例如 NSVF），因为它可以轻松地实时处理数百万个体素**。 我们的实验表明，NSVF 表示中的 10k ∼ 100k 稀疏体素足以对复杂场景进行逼真的渲染。

**Ray Marching inside Voxels.** 我们通过使用等式(2) 沿光线采样点来返回颜色 C(p0, v)。 为了处理光线错过所有对象的情况，我们在等式(2) 的右侧额外添加了一个背景项 。 其中我们定义透明度 ， 是可学习的背景 RGB 值。正如第 2 节中所讨论的，**体渲染需要在非空空间中沿着光线的密集样本来实现高质量渲染**。在整个空间中的均匀采样点进行密集评估（如下图 1 (a)）是低效的，因为经常且不必要地测试空白区域。为了专注于更重要区域的采样，Mildenhall 等人。 (2020) 学习了两个网络，其中第二个网络使用第一个网络估计的分布中的样本进行训练（如下图 1 (b)）。然而，这进一步增加了训练和推理的复杂性。相比之下，**NSVF 不采用二次采样阶段**，同时实现更好的视觉质量。如图 1 (c) 所示，我们使用基于稀疏体素的抛弃采样创建了一组查询点。与上述方法相比，我们能够以相同的评估成本进行更密集的采样。我们包括所有体素交点作为附加样本，并使用中点规则执行颜色累积。我们的方法在算法 1 中进行了总结，其中我们额外返回透明度 A 和预期深度 Z，可进一步用于可视化具有有限差异的法线。





**Early Termination（提前终止）**. NSVF 可以同样好地表示透明和实体对象。 然而，对于实体表面，建议体渲染沿着光线分散表面颜色，这意味着在表面后面需要许多不必要的累积步骤才能使累积透明度达到 0。因此我们使用启发式和当累积透明度  低于某个阈值时，停止评估点。在我们的实验中，我们发现设置 = 0.01 显着加速了渲染过程，而不会导致任何明显的质量下降。

### 学习

由于我们的渲染过程是完全可微的，NSVF 可以通过将渲染的输出结果与一组目标图像进行比较，通过反向传播进行端到端优化，无需任何 3D 监督。 为此，将以下损失最小化：



其中 R 是一批采样光线，C\* 是相机光线的真实颜色，Ω(.) 是 Lombardi(2019) 等人提出的 beta 分布正则化器。 接下来，我们提出了一种渐进式训练策略，以更好地促进学习和推理。

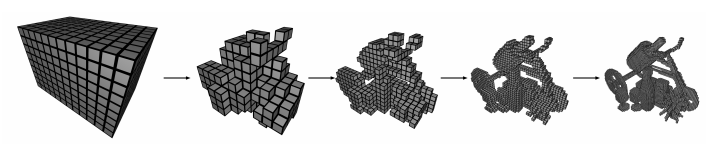
**Voxel Initialization** 我们首先学习用于细分初始边界框（具有体积 V ）的一组初始体素的隐函数，该边界框以足够的间隔粗略包围场景。 初始体素大小设置为 。 如果粗粒度的几何（例如扫描点云或视觉船体输出）可用，初始体素也可以通过对粗粒度的几何进行体素化来初始化。

**Self-Pruning** 现有的基于体积的神经渲染工作（Lombardi 等人，2019 年；Mildenhall 等人，2020 年）表明，**在训练后提取粗略的场景几何图形是可行的**。 基于这一观察，我们提出了一种策略——自剪枝——**一种在训练期间基于粗几何信息有效去除非必要体素的策略，可以使用模型对密度的预测进一步描述**。 也就是说，我们确定要修剪的体素如下**：**



其中 是体素 Vi 内的 G 个均匀采样点（在我们的实验中 G = ），是点 pj 处的预测密度，γ 是一个阈值（在我们所有实验中 γ = 0.5）。 由于此修剪过程不依赖于其他处理模块或输入提示，因此我们将其称为自修剪。 **在粗略的场景几何体出现后，我们定期对体素进行自我修剪**。

**Progressive Training** **上述修剪策略使我们能够逐步调整体素化以适应底层场景结构，并自适应地为重要区域分配计算和内存资源**。 假设学习从初始光线行进步长 τ 和体素大小 l 开始。 经过一定的训练步骤后，我们将 τ 和 l 减半以用于下一阶段。 具体来说，当体素大小减半时，我们将每个体素细分为 个子体素，并且通**过原始八个体素顶点处的特征表示的三线性插值初始化新顶点的特征表示**（即第 3.1 节中的 g~(.)） . 请注意，当使用嵌入作为体素表示时，**我们实质上是逐步增加模型容量以了解场景的更多细节**。 在我们的实验中，我们训练了 4 个阶段的合成场景和 3 个阶段的真实场景。自修剪和渐进式训练的说明如下图 2 所示。



## 实验

我们在多个任务上评估提议的 NSVF，包括多场景学习、动态和大规模室内场景的渲染以及场景编辑和合成。 我们还进行消融研究以验证渐进式训练中不同类型的特征表示和不同选项。 有关架构、实现、数据集预处理和其他结果的更多详细信息，请参阅附录。 另请参阅显示渲染质量的补充视频。

### 实验设置

**Dataset**。

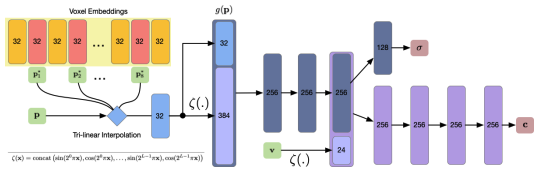
1. Synthetic-NeRF：Mildenhall(2020)等人使用的包括八个物体的合成数据集。
2. Synthetic-NSVF：我们另外渲染了八个具有更复杂几何形状和照明效果的相同分辨率的物体。
3. BlendedMVS：我们在 Yao(2020) 等人数据集的四个物体上进行测试。渲染图像与真实图像混合以获得逼真的环境照明。
4. Tanks & Temples:：我们评估了来自 Knapitsch（2017） 等人的五个物体。我们使用图像并自己标记对象掩码。
5. ScanNet：我们使用 ScanNet 的两个真实场景（Dai 等，2017）。 我们从原始视频中提取 RGB 和深度图像。
6. Maria 序列：这个序列由 Volucap 提供，带有一个在移动的女性的200 帧的网格。 我们渲染每个网格以创建数据集。

**Baselines**。我们采用以下三种最近提出的方法作为基线：

* 场景表示网络（SRN、Sitzmann 等人，2019b）、
* 神经体积（NV、Lombardi 等人，2019 年）
* 神经辐射场（NeRF、Mildenhall 等人） ., 2020)

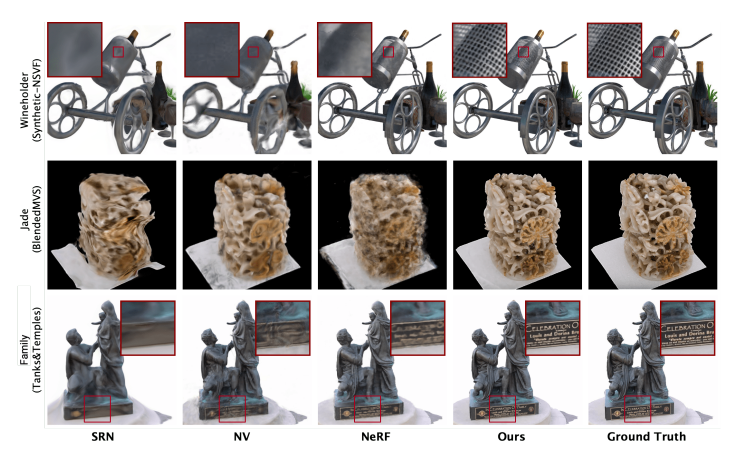
分别代表基于表面的渲染、显式和隐式体积渲染。 有关实现细节，请参阅附录。

**实现细节**。我们使用每个顶点的 32 维可学习体素嵌入对 NSVF 进行建模，并应用（Mildenhall 等人，2020）提出的位置编码。 整体网络架构如下图 9 所示。对于所有场景，我们在 8 个 Nvidia V100 GPU 上使用 32 张图像的批量大小训练 NSVF，并且对于每张图像我们采样 2048 条光线。 为了提高训练效率，我们使用了一种偏置采样策略，只对命中至少一个体素的光线进行采样。 对于所有实验，我们每 2500 步定期修剪体素，并分别在 5k、25k 和 75k 处逐步将体素和步长减半。 我们的开源代码库地址是 <https://github.com/facebookresearch/NSVF>

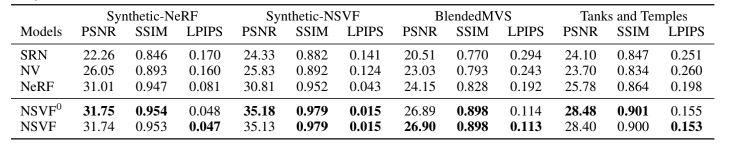


### 结果

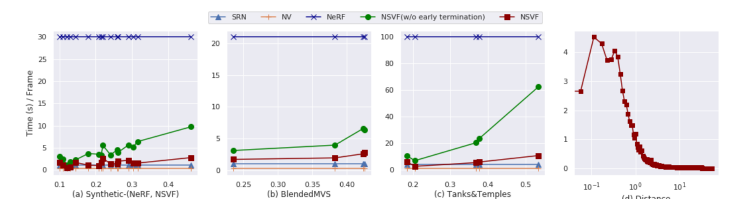
质量比较。我们在下图 3 中展示了定性比较。SRN 倾向于产生过于平滑的渲染和不正确的几何图形； NV 和 NeRF 效果更好，但仍然无法像 NSVF 那样清晰地合成图像。 NSVF 可以在具有复杂几何形状、薄结构和照明效果的各种场景上实现照片般逼真的效果。



此外，如下表 1 所示，NSVF 在所有指标的所有四个数据集上均显着优于三个基线。 请注意，具有提前终止 (= 0.01) 的 NSVF 产生的质量与没有提前终止的 NSVF 几乎相同（在表 1 中表示为 ）。 这表明提前终止不会导致明显的质量下降，同时显着加速计算，如下所示。



速度比较。我们在下图 4 中提供了四个数据集模型的速度比较，其中我们将 Synthetic-NeRF 和 Synthetic NSVF 的结果合并在同一图中，这是因为考虑到它们的图像大小相同。 对于我们的方法，如图 4 (a)- (c) 所示**平均渲染时间与前景与背景的平均比率相关**。这是因为前景的平均比例越高，与体素相交的光线就越多。 因此，需要更多的评估时间。 **平均渲染时间也与相交体素的数量相关**。 当光线在实体对象的渲染中与大量体素相交时，提前终止通过避免表面后面的许多不必要的累积步骤来显着减少渲染时间。 这两个因素可以在图 4 (d) 中看到，我们展示了一个缩小示例。

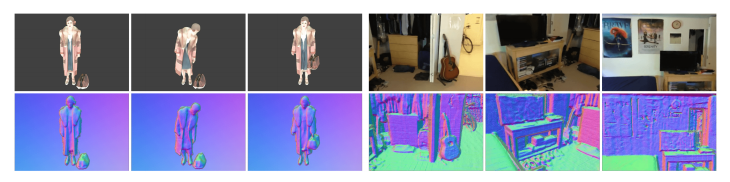


对于其他方法，渲染时间几乎是恒定的。 这是因为他们必须以固定步长评估所有像素，这表明无论光线是否击中场景，无论场景复杂度如何，都会沿每条光线采样固定数量的点。 一般来说，我们的方法比最先进的方法 NeRF 快 10~20 倍，并接近 SRN 和 NV。

存储比较。NSVF 网络权重的存储使用量从 3.2 ∼ 16MB（包括 MLP 的约 2MB）不等，具体取决于所使用的体素数量（10 ∼ 100K）。 NeRF 有两个（粗略和精细）稍深的 MLP，总存储使用量约为 5MB。

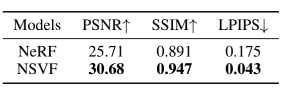
### 对室内场景和动态场景的渲染

我们证明了我们的方法在具有挑战性的由内向外重建场景下对 ScanNet 数据集的有效性。 我们的结果如下图 5 所示，其中初始体素建立在深度图像的点云上。



如上图 5 所示，我们还使用 Maria Se 序列在具有动态场景的语料库上验证了我们的方法。 为了适应 NSVF 的时间序列，我们应用了 Sitzmann(2019b) 等人提出的超网络。 我们还在附录中包含了定量比较，这表明 NSVF 在两种情况下都优于所有基线。

**多场景学习**。我们为来自 Synthetic-NeRF 的所有 8 个物体以及来自 Synthetic-NSVF 的 2 个附加物体（酒架、火车）训练了一个模型。 我们对每个场景使用不同的体素嵌入，但共享相同的 MLP 来预测密度和颜色。为了进行比较，我们基于hypernetwork（Ha 等人，2016 年）在相同的数据集上训练 NeRF 模型。 **如果没有体素嵌入，NeRF 必须使用网络参数对所有场景细节进行编码**，这导致与单个场景学习结果相比质量急剧下降。 表 2 表明我们的方法在多场景学习任务上明显优于 NeRF。



**场景编辑和场景构成**。如下图 6 所示，学习到的多对象模型可以很容易地通过复制和移动体素来组成更复杂的场景，并且可以在没有开销的情况下以相同的方式进行渲染。 此外，我们的方法还通过直接调整稀疏体素的存在来支持场景编辑（参见图 6 中酒架的重新组合）。



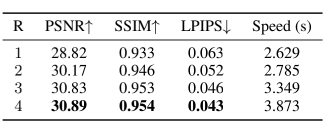
### Albation Studies

我们使用 Synthetic-NSVF 数据集中的一个对象（wineholder），该数据集由具有复杂局部模式（网格）的部分组成，用于消融研究。

**Effect of Voxel Representations(体素表示的影响)**。下图 7 显示了用于编码空间位置的不同类型特征表示的比较（POS 表示使用位置编码，EMB 表示使用体素嵌入）。 **体素嵌入比使用位置编码带来了更大的质量改进**。 此外，通过位置编码和体素嵌入，该模型实现了最佳质量，尤其是在恢复高频模式方面。



**Effect of Progressive Training**。我们还研究了渐进式训练的不同选项（见下表 3）。 请注意，所有模型都仅使用体素嵌入进行训练。 **通过更多轮渐进式训练，性能得到改善。 但是经过一定的轮数后，随着渲染时间的增加，质量只会缓慢提高**。 基于这一观察，我们的模型在实验中进行了 3-4 轮渐进式训练。



## 相关工作

**神经渲染**。最近的工作通过用神经网络替换或增强传统的图形渲染（通常称为神经渲染）显示了令人印象深刻的结果。 我们向读者推荐最近关于神经渲染的调查（Tewari 等人，2020 年；Kato 等人，2020 年）。

* **带有 3D 输入的新视角合成**：DeepBlending（Hedman 等人，2018 年）在几何代理上基于图像的渲染预测混合权重。 其他方法（Thies et al., 2019; Kim et al., 2018; Liu et al., 2019a, 2020; Meshry et al., 2019; Martin Brualla et al., 2018; Aliev et al., 2019）首先将给定的具有显式或神经纹理的几何图形渲染为粗略的 RGB 图像或特征图，然后将其转换为高质量图像。 但是，**这些作品需要 3D 几何体作为输入，并且性能会受到几何体质量的影响**。
* **不带有 3D 输入的新视角合成**：其他方法从 2D 图像学习新视角合成的场景表示。 生成查询网络 (GQN)（Eslami 等人，2018 年）学习 3D 场景的矢量化嵌入，并从新视角中对其进行渲染。 然而，**它们不像 NSVF 那样明确地学习几何场景结构，并且它们的渲染相当粗糙**。 后续工作学习了更多 3D 结构感知表示和伴随的渲染器（Flynn 等人，2016 年；Zhou 等人，2018 年；Mildenhall 等人，2019 年），以多平面图像 (MPI) 作为代理，仅渲染受限制的 插入输入视图的新颖视图范围。 阮福等。 (2018, 2019); 刘等人。 (2019c) 使用基于 CNN 的解码器进行可微渲染以渲染表示为粗粒度体素网格的场景。 然而，由于 2D 卷积核，这种基于 CNN 的解码器无法确保视图一致性。

**神经隐式表示**。已经有相关研究工作使用隐式表示来使用神经网络对 3D 几何进行建模。 与显式表示（如点云、网格、体素）相比，**隐式表示是连续的并且具有高空间分辨率**。 大多数作品在训练期间**需要 3D 监督**来推断任何 3D 点的 SDF 值或占用概率（Michalkiewicz 等人，2019 年；Mescheder 等人，2019 年；Chen 和 Zhang，2019 年；Park 等人，2019 年； Peng 等，2020），而其他作品仅从具有可微渲染器的图像中学习 3D 表示（Liu 等，2019d；Saito 等，2019，2020；Niemeyer 等，2019；Jiang 等，2020 ）。

## 限制和未来的工作

尽管 NSVF 可以有效地生成高质量的新视角并显着优于现有方法，但存在三个主要限制：

(i) 我们的方法无法处理背景复杂的场景。我们假设一个简单的常数背景项 ()。但是，从不同的角度观看真实场景时，通常会有不同的背景。这使得在不干扰目标场景学习的情况下正确捕捉它们的效果变得具有挑战性。

(ii) 我们在所有实验中将自我修剪的阈值设置为 0.5。虽然这对一般场景很有效，但如果阈值设置不正确，对于非常薄的结构可能会发生不正确的修剪。

(iii) 类似于 Sitzmann 等人。 (2019b);米尔登霍尔等人。 (2020)，NSVF 学习颜色和密度作为查询点位置和相机光线方向的“黑箱”函数。因此，渲染性能高度依赖于训练图像的分布，当训练数据不足以预测复杂的几何、材料和光照效果时，可能会产生严重的伪影（见下图 14，玻璃瓶上的折射没有学习）正确）。未来一个可能的方向是将传统的辐射和渲染方程作为物理归纳偏置纳入神经渲染框架。这可以潜在地提高神经网络模型的鲁棒性和泛化能力。



(iv) 当前的学习范式需要已知的相机姿势作为输入来初始化光线及其方向。 对于真实世界的图像，目前没有处理相机校准中不可避免的错误的机制。 当我们的目标数据由多个对象的单视图图像组成时，在实际应用中获得准确配准的姿态就更加困难。 未来研究的一个有前途的途径是使用无监督技术，如 GAN（Nguyen-Phuoc 等人，2019）同时预测相机姿势以获得高质量的自由视角渲染结果。

## 结论

我们提出了 NSVF，一种用于快速和高质量自由视角渲染的混合神经场景表示。 大量实验表明，NSVF 通常比最先进的技术（即 NeRF）快 10 倍以上，同时还能实现更好的质量。 NSVF 可以轻松应用于场景编辑和合成。 我们还演示了各种具有挑战性的任务，包括多场景学习、移动人的自由视角渲染和大规模场景渲染。

## Broader Impact

NSVF 提供了一种从图像中学习神经隐式场景表示的新方法，能够更好地将网络容量分配给场景的相关部分。通过这种方式，它可以比以前的方法更详细地学习大规模场景的表示，这也导致渲染图像的视觉质量更高。此外，所提出的表示比现有技术的渲染速度要快得多，并且可以更方便地进行场景编辑和合成。这种从图像进行 3D 场景建模和渲染的新方法补充并部分改进了已建立的计算机图形概念，并在许多应用中开辟了新的可能性，例如混合现实、视觉效果和计算机视觉任务的训练数据生成。同时，它展示了学习在其他领域具有潜在相关性的空间感知场景表示的新方法，例如目标场景理解、目标识别、机器人导航或基于图像重建的训练数据生成。

仅从 2D 图像中以非常高的视觉保真度捕获和重新渲染真实世界场景模型的能力，也使得在场景中重建和重新渲染人类成为可能。因此，任何对此以及所有相关重建方法的研究和实际应用都必须严格尊重人格权和隐私规定。