神经稀疏体素场

刘凌杰[y]*、顾嘉高[z]*林庆祖[欢]蔡美儿[欢]基督教玄武岩[y]

[y]马克斯·普朗克信息学研究所

[z]脸书人工智能研究[欢]新加坡国立大学

y {lliu, theobalt}@mpi-inf.mpg.德

z jgu@fb.com[欢]kyawzl@comp .nus .edu .sg

[欢]dcscts@nus.edu.sg

摘要

使用经典的计算机图形技术对真实世界的场景进行逼真的自由视点渲染是具有挑战性的，因为它需要捕捉详细的外观和几何模型的困难步骤。最近的研究表明，通过学习在没有3D监督的情况下隐式地编码几何和外观，有希望的结果。然而，现有的方法在实践中往往由于有限的网络容量或难以找到相机光线与场景几何图形的准确交叉点而显示出模糊的渲染。从这些表示法中合成高分辨率的图像通常需要耗时的光学光线行进。在这项工作中，我们引入了神经稀疏体素场（NSVF），一种新的神经场景表示，用于快速和高质量的自由视点渲染。NSVF定义了一组以稀疏体素八叉树组织的体素边界隐式字段，以建模每个单元格中的局部属性。我们仅从一组假设的RGB图像中，通过一个可扩散的射线行进操作，逐步学习底层的体素结构。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新视图的呈现。我们的方法通常在推理时间上比最先进的（即NeRF（Mildenhall等人，2020））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，我们的方法可以很容易地应用于场景编辑和场景合成。我们还演示了几个具有挑战性的任务，包括多场景学习、一个移动的人类的自由视点渲染、和大规模的场景渲染。

1介绍

计算机图形学中的现实渲染具有广泛的应用范围，包括混合现实、视觉效果、可视化，甚至是计算机视觉和机器人导航中的训练数据生成。从任意的角度逼真地渲染现实世界的场景是一个巨大的挑战，因为获得高质量的场景几何形状和材料模型通常是不可行的，就像在高预算的视觉效果制作中所做的那样。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点插值相结合（Shum和Kang，2000；张和陈，2004；泽利斯基，2010）。尽管IBR方法取得了重大进展，但该方法仍然具有次优的渲染质量和对结果的控制有限，而且通常是特定于场景类型的。为了克服这些限制，最近的研究已经使用了深度神经网络来隐式地学习包含几何和外观的场景表示。这种神经表征通常与三维几何模型相结合，如体素网格(Yan等人，2016；Sitzmann等人，2019a；

*同样的贡献。

第34届神经信息处理系统会议（NeurIPS 2020），加拿大温哥华。

隆巴迪等人，2019年），纹理网格（蒂斯等人，2019年；金等人，2018年；刘等，2019a，2020年）、多平面图像（周等人，2018年；弗林等人，2019年；米尔登霍尔等，2019年）、点云（梅氏等，2019年；阿利耶夫等，2019年）和隐式函数（西茨曼等，2019b；米尔登霍尔等，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，在理论上可以实现高空间分辨率。然而，现有的方法在实践中往往由于有限的网络容量或难以找到相机光线与场景几何图形的准确交叉点而显示出模糊的渲染。从这些表示法中合成高分辨率的图像通常需要耗时的光学光线行进。此外，用这些神经表征来编辑或重新合成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），一种新的快速和高质量的自由视点渲染的隐式表示方法。NSVF不是用单个隐式函数来建模整个空间，而是由一组由稀疏体素八叉树组织的体素边界隐式字段组成。具体来说，我们在体素的每个顶点上分配一个体素嵌入，并通过在相应体素的8个顶点上聚合体素嵌入，获得体素内查询点的表示。这将进一步通过一个多层感知器网络（MLP）来预测该查询点的几何形状和外观。我们的方法可以仅从一组场景的二维姿态图像中，通过可区分的射线行进操作，从粗到细逐步学习NSVF。在训练过程中，对不包含场景信息的稀疏体素进行修剪，使网络专注于具有场景内容的体积区域的隐式函数学习。使用稀疏体素，通过跳过没有场景内容的空体素，可以大大加速推理时的渲染。

我们的方法通常在推理时间上比最先进的（即NeRF（Mildenhall等人，2020））快10倍以上，同时获得更高质量的结果。我们广泛地评估了我们在各种具有挑战性的任务上的方法，包括多对象学习，动态场景和室内场景的自由视点渲染。我们的方法可以用于编辑和合成场景。总之，我们的技术贡献是：

- o我们提出了由一组体素有界的隐式字段组成的NSVF，其中对于每个体素，体素嵌入被学习来编码局部属性，以进行高质量的渲染；
- o NSVF利用稀疏体素结构来实现高效的渲染；
- o我们引入了一种渐进的训练策略，该策略通过可微的光线行进操作，从一组姿态化的二维图像中以端到端的方式有效地学习底层的稀疏体素结构。

2背景

现有的神经场景表示和神经渲染方法通常旨在学习一个函数，空间位置映射到一个特征表示，隐式地描述当地的几何形状和外观的场景，小说场景的视图可以合成使用渲染技术。为此，渲染过程以可微的方式表述，通过最小化渲染和二维图像之间的差异来训练编码场景表示的神经网络。在本节中，我们将描述使用隐式字段进行表示和渲染的现有方法及其局限性。

.12具有隐式字段的神经渲染

让我们用一个隐式函数F来表示一个场景e：（p，v）!θ(c，!),其中是一个底层神经网络的参数。这个函数描述了场景颜色c及其概率密度！在空间位置p和射线方向v。给定在p位置有一个针孔照相机02 $R^{3}\times$，我们通过从相机拍摄光线到3D场景来渲染大小为HW的2D图像。因此，我们评估了一个体积渲染积分来计算相机射线p（z）= p的颜色0+z·v为：

$$\boldsymbol{C}(\boldsymbol{p}_0,\boldsymbol{v})=\int_0^{+\infty}\omega(\boldsymbol{p}(z))\cdot\boldsymbol{c}(\boldsymbol{p}(z),\boldsymbol{v})dz,\quad \text{where}\int_0^{+\infty}\omega(\boldsymbol{p}(z))dz=1 \tag{1}$$

请注意，为了鼓励场景表示是多视图一致的，！被限制为仅为p（z)的函数，而c同时接受p（z）和v作为模型与视图相关的颜色的输入。采用不同的渲染策略来评估这个积分是可行的。

表面渲染。基于表面的方法（Sitzmann等人，2019b；Liu等人，2019b；Niemeyer等人，2019年）假设！（p（z））为Dirac函数6（p(z)-p(z)），其中p（z)是相机光线与场景几何的交集。

卷渲染。基于体积的方法（Lombardi等人，2019年；Mildenhall等人，2020年）估计积分C（p0，v)在等式1通过在每个相机射线上密集采样点，并将采样点的颜色和密度累积成二维图像。例如，最先进的方法NeRF（Mildenhall等人，2020年）估计了C（p0，v）为：

$$\boldsymbol{C}(\boldsymbol{p}_0, \boldsymbol{v}) \approx \sum_{i=1}^{N} \left(\prod_{j=1}^{i-1} \alpha(z_j, \Delta_j) \right) \cdot (1 - \alpha(z_i, \Delta_i)) \cdot \boldsymbol{c}(\boldsymbol{p}(z_i), \boldsymbol{v}) \tag{2}$$

其中α（zi,Δi)=exp(a(p(zi）·Δi和Δi= zi+1— zi. {c(p(zi$_{i=1}^{N}$)，v)}和{a（p(zi$_{i=1}^{N}$是采样点的颜色和体积密度。

.22.现有方法的局限性

对于表面渲染，找到一个精确的表面对于学习到的颜色是多视图一致的是至关重要的，这不利于训练的收敛，从而在渲染中导致模糊。体渲染方法需要沿光线对大量点进行采样以积累颜色，以实现高质量的渲染。然而，像NeRF一样评估沿射线的每个样本点是低效的。例如，NeRF渲染一个800年800年的图像大约需要30秒。×我们的主要观点是，在尽可能没有相关场景内容的情况下，防止对空白空间中的点进行采样是很重要的。虽然NeRF沿着射线进行重要的采样，但由于为每条射线分配了固定的计算预算，它不能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，如边界体积层次（BVH，Rubin和Whitten，1980）和稀疏体素八叉树（SVO，Laine和Karras，2010），它们被设计用来在稀疏分层结构中建模场景，用于光线跟踪加速。在这种编码中，一个空间位置的局部属性只依赖于该空间位置所属的叶节点的一个局部邻域。在本文中，我们展示了如何将分层稀疏体积表示用于三维场景的神经网络编码的隐式领域，以实现详细的编码，以及高效、高质量的可微体积渲染，甚至是大规模的场景。

3个神经稀疏体素场

在本节中，我们将介绍神经稀疏体素场（NSVF），这是一种结合了神经隐式域和显式稀疏体素结构的混合场景表示。NSVF不是将整个场景表示为一个单一的隐式字段，而是由一组以稀疏体素八叉树组织的体素边界隐式字段组成。下面，我们描述NSVF的构建块-一个体素有边界的隐式字段(3。1) -然后是一个NSVF的渲染算法（3.2），以及一个渐进式学习策略（3.3）。

.13个体素有界的隐式字段

我们假设一个场景中相关的非空部分包含在一组稀疏（边界）体素V = {V中1. ..VK，并且场景被建模为一组体素有界的隐式函数： F（p，v）= F（g$_{\theta}^{i}$i(p)、v）如果是p 2 Vi$_{\theta}^{i}$.每个F都被建模为一个具有共享参数的多层感知器（MLP）：θ

$$F : (g_{\theta}^{i}i(p),\ v) \ !\ (c、a)、Ap\ 2\ V_i, \tag{3}$$

这里c和a是三维点p的颜色和密度，v是射线方向，gi(p)为在p处的表示法，其定义为：

$$gi\zeta(p) = (x\ (\widetilde{g}\,i_1^*(p), \ldots, \widetilde{g}\,i_8^*(p))) \tag{4}$$

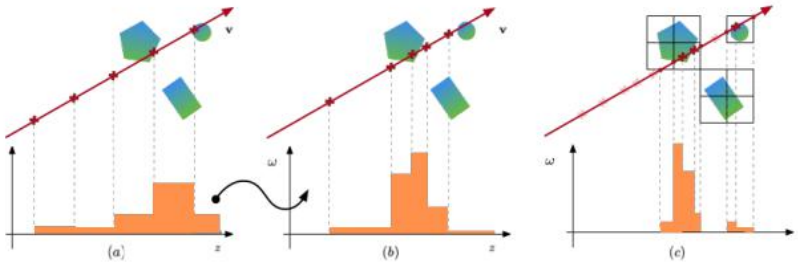

图1：(a)均匀抽样示意图；基于(a)中结果的(b)重要性抽样；(c)提出了一种基于稀疏体素的采样方法。

p的位置。$\tilde{g}_1^*..$，p 2 R_8^{*3}是V的八个顶点吗i和$\tilde{g}_{i1}^*$(p)，...，$\tilde{g}$i(p) 2 R_8^{*d}是存储在每个顶点上的特征向量。此外，x（.）指三线性插值，和（.）是一个后处理功能。ζ在我们的实验中，（.）是由（Vaswani等人提出的位置编码，2017；米尔登霍尔等人，2020年）。ζ

与使用点p的三维坐标作为F的输入相比$_e^i$正如之前的大部分工作一样，在NSVF中，特征表示gi(p)由相应体素的8个体素嵌入聚合，其中区域特定的信息(e。g.几何形状，材料，颜色)都可以嵌入。它显著地简化了后续F的学习$_e^i$以及便于高质量的渲染。

特殊情况。NSVF包含了两类早期的作品作为特殊情况。(1)何时$\tilde{g}$i$_{\tilde{g}\tilde{g}}^{**}$ζ(p) = p和（.）是位置编码，giζ(p) = （x (p , .$\tilde{g}_1^*..\tilde{g}_8^*$ζ, p)) = (p)，这意味着NeRF（米尔登霍尔等人，2020年）是NSVF的一个特例。(2)何时$\tilde{g}$i(p) : p !ζ(c, a)，(.)和“F”$_e^i$如果是身份函数，我们的模型等价于使用显式体素来存储颜色和密度的模型，例如，神经体积（Lombardi et al.，2019）。

3.2卷渲染

NSVF编码一个场景的颜色和密度。与渲染对整个空间建模的神经隐式表示相比，渲染NSVF更有效，因为它排除了空白空间中的采样点。渲染分两个步骤执行：(1)射线体素交叉点；以及(2)体素内部的射线行进。我们在附录图8中说明了这个管道，

射线体素交叉点。我们首先对每条射线应用轴对齐边界盒相交检验（AABBtest）（Haines，1989）。它通过比较从射线原点到体素的六个边界平面的距离来检查射线是否与体素相交。AABB测试非常有效，特别是对于层次八叉树结构(e。g.NSVF)，因为它可以很容易地实时处理数百万体素。~我们的实验表明，NSVF表示中的10k 100k稀疏体素足以实现复杂场景的逼真渲染。

射线标记在体素内。我们返回颜色C (p0，v)通过使用等式沿着射线进行采样点(2).为了处理射线错过所有对象的情况，我们另外添加了一个背景项

A(p0, v) • cbg在等式的右边*(2)，其中我们定义了透明度A (p*0, v) = $^{u_N}_{i=1}$ α (zi,Δi)，和cbg是可学习的背景RGB值。如2中所述，体渲染需要在非空空间沿射线密集样本才能实现高质量的渲染。在整个空间的均匀采样点进行密集评估（图1 (a)）是低效的，因为空区域被频繁和不必要的测试。为了关注在更重要的地区的抽样，米尔登霍尔等人。（2020）学习了两个网络，其中第二个网络是由第一个网络估计的分布样本进行训练的（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没有采用二次采样阶段，同时获得更好的视觉质量。如图1 (c)所示，我们使用基于稀疏体素的拒绝采样创建了一组查询点。与上述方法相比，我们能够在相同的评估成本下更密集地抽样。我们将所有体素交点作为额外的样本，并使用中点规则执行颜色积累。我们的方法在算法1中总结，我们另外返回透明度A和期望深度Z，期望深度Z可以进一步用于有限差分的法态可视化。

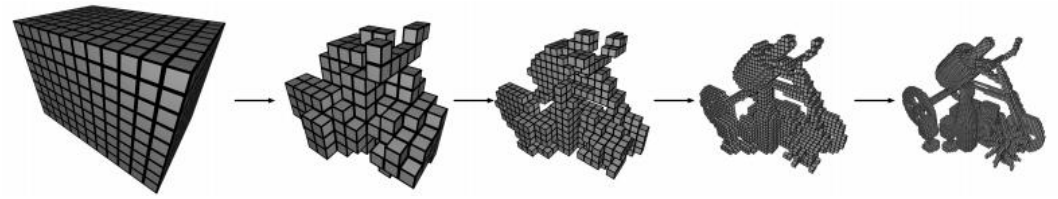

图2：自修剪和渐进式训练的说明

提前终止。NSVF可以同样很好地表示透明物体和固体物体。然而，对于固体表面，所提出的体积渲染沿着射线分散了表面的颜色，这意味着它需要在表面后面进行许多不必要的积累步骤，才能使积累的透明度A (p0，v)达到0。因此，我们使用启发式，并停止计算点时，积累的透明度A (p0，v)低于某一阈值e。在我们的实验中，我们发现设置e = 0.01显著加速了渲染过程，而没有导致任何明显的质量下降。

.33学习

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

$$\mathcal{L} = \sum_{(\boldsymbol{p}_0,\boldsymbol{v})\in R} \|\boldsymbol{C}(\boldsymbol{p}_0,\boldsymbol{v}) - \boldsymbol{C}^*(\boldsymbol{p}_0,\boldsymbol{v})\|_2^2 + \lambda \cdot \Omega\left(A(\boldsymbol{p}_0,\boldsymbol{v})\right), \tag{5}$$

*其中R是一批采样的射线，C是相机射线的地面真实颜色，和（。）是Lombardi等人提出的一种beta分布正则化器。(2019).接下来，我们提出了一个渐进的训练策略，以更好地促进学习和推理：

体素初始化我们首先学习初始体素集细分初始边界框的隐式函数_V)，以足够的空间大致包围了场景。初始体素大小设置为l≈ $\sqrt[3]{}$V/1000.如果是一个粗糙的几何图形(e。g.扫描点云或视觉船体输出)是可用的，初始体素也可以初始化体素初始化粗几何。

自修剪现有的基于体积的神经渲染工作（Lombardi等人，2019；米尔登霍尔等人，2020）已经表明，在训练后在粗糙水平上提取场景几何是可行的。基于这一观察结果，我们提出了一种-自剪枝——在基于粗几何信息的训练过程中有效去除非必要体素的策略，该策略可以通过模型的密度预测来进一步描述。也就是说，我们确定要修剪的体素如下：

$$\text{Vi如果最小的} \exp\left(-\sum_{j=1\ldots G} a(g_i(p_j))\right) > V, \quad p_j \in V_i,\ V_i \in V, \tag{6}$$

其中{pj$_{j=1}^{G}$G是在体素V内的均匀采样点吗i(G = 16^{3}在我们的实验中)i(pj))为p点的预测密度j，V是一个阈值（在我们所有的实验中，V=为0.5）。由于这个修剪过程不依赖于其他处理模块或输入线索，我们称之为自修剪。在粗糙的场景几何图形出现后，我们周期性地对体素进行自剪枝。

上述剪枝策略使我们能够逐步调整体素化到底层场景结构，并自适应地将计算和内存资源分配到重要区域。假设学习从初始的射线行进步长τ和体素大小l开始。经过一定步骤的训练后，我们将τ和l减半到下一阶段。具体来说，当将体素大小减半时，我们将每个体素细分为2^{3}子体素和新顶点的特征表示。e.$\tilde{g}$(.)在Subar3。1)通过原始8个体素顶点的特征表示的三线性插值初始化。注意，当使用嵌入作为体素表示时，我们本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，我们用4个阶段训练合成场景，用3个阶段训练真实场景。图2显示了自修剪和渐进式训练的说明。

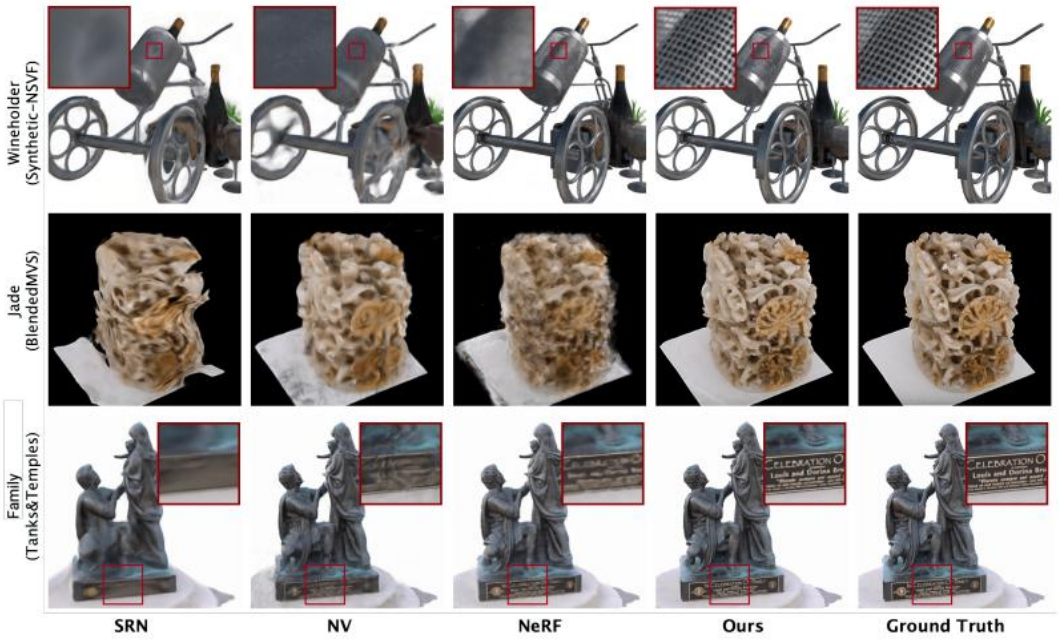

图3：对来自单个场景数据集的场景的测试视图的比较。对于获胜者和家庭，特写镜头显示为更清晰的视觉比较。

4实验

我们在多场景学习、动态和大规模室内场景的渲染、场景编辑和合成等任务上评估了所提出的NSVF。我们还进行了消融研究，以验证不同类型的特征表征和渐进式训练中的不同选择。有关体系结构、实现、数据集的预处理和其他结果的更多细节，请参阅附录。也请参考显示渲染质量的补充视频。

4.1实验设置

数据集(1)合成-nerf：米尔登霍尔等人使用的合成数据集。（2020）包括8个对象。*(2)合成-NSVF：我们另外渲染8个对象在相同的分辨率与更复杂的几何形状和照明效果。(3)我们测试了姚等人的四个物体。*(2020).渲染的图像与真实的图像混合，以具有真实的环境照明。(4)坦克和寺庙：我们对克纳皮奇等人的五个物体进行了评估。（2017），我们使用图像和标签的对象掩盖自己。*(5) ScanNet：我们使用了ScanNet中的两个真实场景（Dai等人，2017）*。我们从原始视频中提取了RGB图像和深度图像。(6)玛丽亚序列：这个序列由Volucap提供，有200帧的移动女性的网格。我们渲染每个网格来创建一个数据集。

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，西茨曼等人，2019年b）、神经体积（NV，Lombardi等人，2019年）和神经辐射场（NeRF，米尔登霍尔等人，2020年），分别表示基于表面的渲染、显式和隐式体积渲染。实施细节见附录。

我们对每个顶点使用32维可学习体素嵌入对NSVF进行建模，然后应用位置编码（米尔登霍尔等人，2020）。整个网络架构如图附录图9所示。对于所有场景，我们在8个NvidiaV100gpu上使用32张批量图像来训练NSVF，对于每个图像，我们采样2048条射线。为了提高训练效率，我们使用了一种有偏的采样策略，只对击中至少一个体素的光线进行采样。在所有的实验中，我们每2500步定期修剪体素，并分别在5k、25k和75k时逐步将体素和步长减半。我们有开源的代码库https://github.com/facebookresearch/NSVF

表1：四个数据集的测试集的定量比较。我们使用三个指标：PSNR（“）、SSIM（”）和LPIPS (t)（Zhang et al.，2018）来评估渲染质量。分数在所有场景的测试图像上取平均值，我们在附录中显示每个场景的分解结果。默认情况下，NSVF是通过提前终止(e = 0。01).我们还显示了没有使用早期终止（e = 0）的结果，表示为NSVF^0.

模型	合成-NeRF PSNR SSIM LPIPS			合成版-NSVF PSNR SSIM LPIPS			请PSNR SSIM LPIPS			坦克和寺庙，PSNR SSIM LPIPS		
合格护士	22.26	0.846	0.170	24.33	0.882	0.141	20.51	0.770	0.294	24..84710	0	0.251
十亿分之一伏	26.05	0.893	0.160	25.83	0.892	0.124	23.03	0.793	0.243	23.70	0.834	0.260
NeRF	.0131	0.947	0.081	.8130	0.952	0.043	24.15	0.828	0.192	25.78	0.864	0.198
NSVF^{0}N SVF	**31.75**	**0.954**	0.048	**35.18**	**0.979**	**0.015**	26.89	**0.898**	0.114	**28.48**	**0.901**	0.155
	31.74	0.953	**0.047**	35.13	**0.979**	**0.015**	**26.90**	**0.898**	**0.113**	28.40	0.900	**0.153**

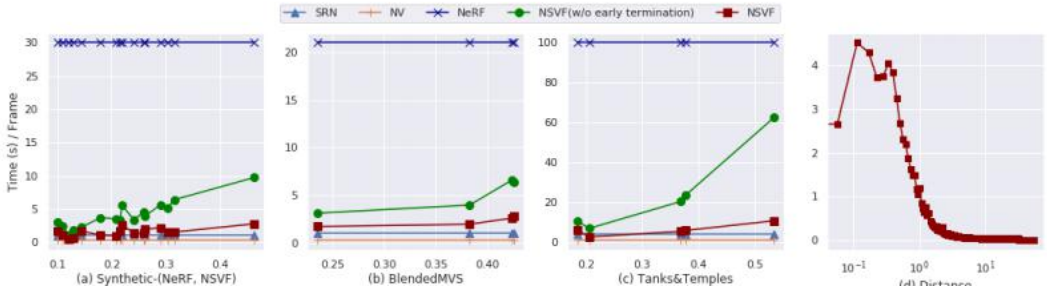

图4：我们报告了为(a)-(c)中的所有数据集渲染一幅图像所花费的时间，其中x轴表示前景与背景的上升比，y轴表示秒内的渲染时间。我们还显示了NSVF在(d)合成场景上渲染时间的曲线曲线，其中x轴表示相机到物体中心的距离，y轴表示秒的渲染时间。

.24结果

质量比较我们在图3中显示了定性比较。SRN倾向于产生过于平滑的渲染和不正确的几何形状；NV和NeRF工作得更好，但仍然不能像NSVF那样清晰地合成图像。NSVF可以在具有复杂几何形状、薄结构和灯光效果的各种场景上实现逼真的效果。

此外，如表1所示，NSVF在所有指标上的所有四个数据集上都显著优于这三个基线。值得注意的是，有早期终止的NSVF（e=0.01）产生的质量几乎与没有早期终止的NSVF相同(表示为NSVF^{0}在表1中）。这表明，早期终止不会导致显著的质量下降，同时显著加速计算，接下来将看到。

*速度比较我们在图4中提供了四个数据集上的模型上的速度比较，其中我们将合成-nerf和合成-NSVF的结果合并在同一图中，考虑到它们的图像大小是相同的。*对于我们的方法，平均渲染时间与前景与背景的平均比值相关，如图4 (a)-(c).所示这是因为前景的平均比例越高，光线与体素相交的次数就越多。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。当一条射线在实体对象的渲染中与大量体素相交时，早期终止通过避免表面后面许多不必要的积累步骤，显著减少了渲染时间。这两个因素可以在图4 (d)中看到，其中我们展示了一个缩小的示例。

对于其他方法，渲染时间几乎是恒定的。这是因为它们必须用固定的步骤来计算所有的像素，这表明无论光线是否到达场景，无论场景的复杂性如何，都要沿着每条光线采样固定数量的点。总的来说，我们的方法比最先进的方法NeRF大约快10 20倍，并且接近于SRN和NV。∼

存储比较NSVF的网络权重的存储使用量从3.2不等∼

16MB（包括mlp的约2mb），这取决于所使用的体素数量（10 100K）。∼NeRF有两个（粗和细）稍深的mlp，总存储使用率约为5MB。

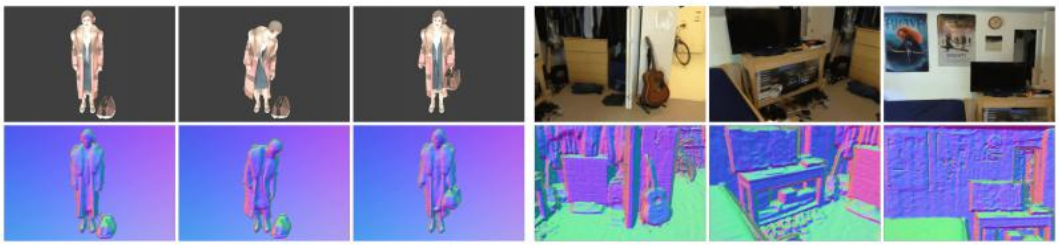

*图5：我们对Maria序列（左）和ScanNet（右）的结果。*我们渲染测试轨迹，并在RGB（上）和相应的表面法线（下）中显示三个采样帧。

模型	信号-噪音功率比	ssim”	LPIPSt
NSVF	**32.04**	**0.965**	**0.020**
w/o POS	30.89	0.954	0.043
w/o EMB	27.02	0.931	0.077
w/o POS, EMB	24.47	0.906	0.118

NSVF w/o EMB NSVF w/o POS NSVF地面真相

图7：左表显示了没有位置编码（没有POS）、没有体素嵌入（没有EMB）或没有两者（没有POS、EMB）的NSVF的定量比较。右边的图显示了与地面真实图像的视觉比较。

图6：合成数据集上的场景合成和编辑。真实的图像显示在右下角。

室内场景和动态场景的渲染，我们展示了我们的有效性
在扫描网数据集下的方法
挑战性由内到外的重建场景。我们的结果如图5所示，其中初始体素是建立在从深度图像中获得的点云之上的。

如图5所示，我们还使用Maria序列在一个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

应用席茨曼等人提出的超网络。(2019b).我们还在附录中包括了定量比较，这表明NSVF在这两种情况下都优于所有基线。

多场景学习，我们训练了一个单一的模型*所有的8个对象从合成-nerf和2个addi-来自合成-NSVF的特定对象（获胜者，火车）。*我们对每个场景使用不同的体素嵌入
共享相同的mlp来预测密度和颜色。为了进行比较，我们训练了同样的NeRF模型基于超网络的数据集（Ha等人，2016）。

表2：多场景学习的结果。

模型	信号-噪音功率比	ssim”	LPIPSt
NeRF	25.71	0.891	0.175
NSVF	**30.68**	**0.947**	**0.043**

在没有体素嵌入的情况下，NeRF必须用网络参数对所有场景细节进行编码，与单个场景学习结果相比，这导致质量急剧下降。表2显示，我们的方法在多场景学习任务上显著优于NeRF。

场景编辑和场景合成。如图6所示，学习到的多对象模型可以很容易地使用，通过复制和移动体素来组成更复杂的场景，并以同样的方式呈现，而没有开销。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中的获胜器的重新组合）。

表3：渐进训练消融。

R	信号-噪音功率比	ssim”	LPIPSt	速度(s)
1	28.82	0.933	0.063	2.629
2	30.17	0.946	0.052	2.785
3	30.83	0.953	0.046	3.349

4	**30.89**	**0.954**	**0.043**	3.873

.34消融研究

我们使用来自合成NSVF数据集的一个对象（获胜者），该对象由具有复杂局部模式（网格）的部分组成，以进行消融研究。

体素表示的效果图7显示了对编码空间位置的不同类型的特征表示的比较。体素嵌入比使用位置编码更能更好地提高质量。此外，通过位置编码和体素嵌入，该模型获得了最好的质量，特别是在恢复高频模式方面。

渐进式训练的效果，我们也研究了渐进式训练的不同选择（见表3）。请注意，所有的模型都只使用体素嵌入进行训练。通过更多回合的渐进式训练，成绩得到了提高。但经过一定的回合数，质量提高，而渲染时间增加。基于这一观察结果，我们的模型在实验中进行了3-4轮的渐进式训练。

5相关工作

神经渲染最近的工作显示了令人印象深刻的结果，通过用神经网络取代或增强传统的图形渲染，这通常被称为神经渲染。我们建议读者参考最近的神经渲染调查（Tewari等人，2020年；Kato等人，2020年）。

o带有3D输入的新视图合成：深度混合（Hedman et al.，2018）在几何代理上预测基于图像的渲染的混合权重。其他方法（Thies等人，2019年；Kim等人，2018年；刘等人，2019年，2020年；Meshry等人，2019年；马丁布鲁拉等人，2018年；Aliev等人，2019）首先将给定的具有显式或神经纹理的几何图形渲染成粗糙的RGB图像或特征图，然后将其转换成高质量的图像。然而，这些工作需要三维几何作为输入，性能将受到几何质量的影响。

o没有3D输入的新视图合成：其他方法从二维图像中学习新视图合成的场景表示。生成查询网络（GQN）（Eslami等人，2018）学习一个三维场景的向量化嵌入，并从新的视图渲染它。然而，它们并不像NSVF那样显式地学习几何场景结构，而且它们的渲染图相当粗糙。后续的工作学习了更多的3d结构感知表示和伴随的渲染器（Flynn等人，2016；周等人，2018；米尔登霍尔等人，2019年），以多平面图像（MPIs）作为代理，这只呈现有限范围的新视图插值输入视图。Nguyen-Phuoc等人。（2018、2019）；刘等人。（2019c）使用基于cnn的解码器进行可微渲染，以渲染以粗粒度体素网格表示的场景。然而，由于二维卷积核的存在，这种基于cnn的解释并不能确保视图的一致性。

神经隐式表示。隐式表示已经被研究用于模型的三维几何与神经网络。与显式表示（如点云、网格、体素）相比，隐式表示是连续的，具有较高的空间分辨率。大多数作品在培训期间需要3D监督，以推断SDF值或任何3D点的占用概率（米哈凯维奇等人，2019；梅舍德等人，2019；陈和张，2019；公园等人，2019；彭等人，2020年），而其他作品仅从具有可微渲染器的图像中学习3D表示(Liu等人，2019年d；齐藤等人，2019年，2020年；Niemeyer等人，2019年；江等人，2020年）。

6结论

我们提出了NSVF，一种用于快速和高质量的自由视点渲染。大量的实验表明，NSVF通常比最先进的技术（即NeRF）快10倍以上，同时获得了更好的质量。NSVF可以很容易地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、移动人类的自由视点渲染和大规模场景渲染。

7更广泛的影响

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到场景的相关部分。通过这种方式，它能够以比以前的方法更高的细节学习大规模场景的表示，这也导致了渲染图像的更高的视觉质量。此外，所提出的表示方式使比最先进的渲染速度更快的渲染，并使更方便的场景编辑和合成。这种从图像中进行三维场景建模和渲染的新方法补充和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，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，它展示了学习其他领域潜在相关性的空间感知场景表示的新方法，如物体场景理解、物体识别、机器人导航或基于图像重建的训练数据生成。

只有从2D图像中，才能捕捉和重新渲染真实世界场景的模型，这也使在一个场景中重建和重新渲染人类成为可能。因此，任何对此和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利和隐私规定。

对资金的确认和披露

我们感谢沃卢卡普·巴贝尔斯伯格和弗劳恩霍夫·海因里希·赫兹研究所提供的玛丽亚数据集。我们也感谢李世伟、陈能伦、本米尔登霍尔对实验的帮助；进行讨论。克里斯蒂安·西奥堡特得到了ERC合并者拨款770784的支持。刘灵杰获得梅特纳博士后资助。本文的计算工作部分是在新加坡国家超级计算中心(https://www.nscc.sg).

参考文献

卡拉阿里阿列耶夫，德米特里尤利亚诺夫和维克多兰皮茨基。2019.神经点图形。*arXiv预印arXiv：1906.08240。*

Z.陈和H。张。2019.学习生成式形状建模的隐式领域。*2019年IEEE/CVF计算机视觉和模式识别（CVPR）会议，第5932-5941页。*

黛安娜，天使X。张，马诺利斯·萨瓦，马切杰·哈尔伯，托马斯·芬克豪瑟和马提亚斯·Nießner。2017.注释丰富的室内场景的三维重建。*在程序中。计算机视觉与模式识别（CVPR），IEEE。*

阿里·埃斯拉米、丹尼洛·希门尼斯·雷曾德、弗雷德里克·贝斯、法比奥·维奥拉、阿里·莫科斯、玛尔塔·加内洛、阿夫拉汉姆·鲁德曼、安德烈·鲁苏、伊沃·丹尼赫尔卡、卡罗尔·格雷戈尔等。2018.神经场景的表示和渲染。科学，360（6394）：1204-1210。

约翰·弗林、迈克尔·布罗克斯顿、保罗·德贝维克、马修·杜瓦尔、格雷厄姆·菲夫、瑞安·奥弗贝克、诺亚·斯纳弗利和理查德·塔克。2019.深度视图：查看合成与学习到的梯度下降。*计算机视觉和模式识别国际会议（CVPR）。*

约翰·弗林，伊凡·纽兰德，詹姆斯·菲尔宾和诺亚·斯纳弗利。2016.深度立体声：学习预测来自世界意象的新视角。*在计算机视觉和模式识别（CVPR）中。*大卫哈，戴安德烈，和

学报。2016.超网络。

埃里克海恩斯。1989.*基本的光线追踪算法，第33-77页。*学术出版社有限公司。，GBR.

彼得·海德曼，朱利安·菲利普，真正的普赖斯，简-迈克尔·弗拉姆，乔治·德雷塔基斯和加布里埃尔·布罗斯托。

2018.深度混合的自由视点基于图像的渲染。*ACM跨。图，* 37(6):257:1 - 257:15.

岳江、丹济、志忠汉、茨威克。2020.Sdfdiff：对三维形状优化的有符号距离字段的可微渲染。*在IEEE/CVF计算机视觉和模式识别（CVPR）会议上。*

加藤广春、贝克、森海、安藤高弘、松冈、凯尔和盖登。2020.可区分的渲染：一种调查。*arXiv预印arXiv：2006.12057*。

金庆宇、巴勃罗·加里多、特瓦里、徐伟鹏、马斯提斯、Nießner、帕特里克·佩雷斯、克里斯蒂安·理查特、迈克尔·佐洛弗和克里斯蒂安·西奥巴特。2018.深视频肖像。*ACM图形事务处理（TOG），37*。

国王、朴杰熙、周千毅、科尔敦。2017.坦克和寺庙：作为大规模场景重建的基准测试。*ACM图形交易，36(4)*。

萨穆利·莱恩和特罗·卡拉斯。2010.高效的稀疏体素八叉树-分析、扩展和实现。

刘灵杰、徐伟鹏、哈伯曼、迈克尔、弗罗里安、伯纳德、金庆宇、王文平、西奥华。2020.通过学习动态纹理和渲染到视频翻译的神经人体视频渲染。*IEEE《可视化与计算机图形学学报》，PP：1-1*。

刘灵杰、徐伟鹏、佐尔霍弗、金庆宇、弗洛里安伯纳德、王文平、西奥华。2019a.人类演员视频的神经渲染和再现。*ACM图形交易（TOG）*。

刘绍辉、张银达、彭松友、石波信、波尔辉、崔兆鹏。2019b.区域：渲染具有可微球面跟踪的深隐式有符号距离函数。*arXiv预印arXiv：1911.13225*。

刘世臣、陈伟凯、李天业、郝浩。2019c.软栅格化器：用于无监督的单视图网格重建的可微渲染。*arXiv预印arXiv：1901.05567*。

刘世臣、齐藤顺助、陈伟凯、李浩。2019d.学习在没有三维监督的情况下推断内隐表面。*《神经信息处理系统的进展》，第8295-8306页*。

斯蒂芬·隆巴迪、托马斯·西蒙、杰森·萨拉吉、加布里埃尔·施瓦茨、安德烈亚斯·莱尔曼和耶泽·谢赫。2019.神经体积：从图像中学习动态可渲染的体积。*ACM图形交易（TOG），38(4)：65*。

里卡多·马丁·布鲁拉、彼得·林肯、阿达什·科德尔、克里斯托夫·雷曼、丹、丹·戈德曼、凯斯金、史蒂夫·塞茨、沙拉姆·伊扎迪、肖恩·法内罗、罗希特·潘迪、杨、帕维尔·皮德利彭斯基、乔纳森·泰勒、朱利安·瓦伦丁、萨米·哈米斯、菲利普·戴维森和阿纳斯塔西娅·特卡奇。2018.外观定位：通过实时神经重新渲染来增强性能捕获。卷37。

拉尔斯·梅切德、迈克尔·奥克斯尔、迈克尔·尼迈耶、塞巴斯蒂安·诺沃津和安德烈亚斯·盖格。2019.占用网络：在功能空间中学习三维重建。*在诉讼IEEE Conf。计算机视觉与模式识别（CVPR）的研究*。

穆斯塔法·梅什里，丹·B。戈德曼，哈米斯，胡格斯霍普，罗希特潘迪，诺亚斯纳维利，和里卡多马丁-布鲁阿拉。2019.在野外的神经再现。*在计算机视觉和模式识别（CVPR）中*。

米哈尔凯维奇，Jhony K。庞特斯，多米尼克·杰克，巴克塔什莫特拉格和安德斯·埃里克森。2019.神经网络中隐含的表面表示。*在IEEE国际计算机视觉会议（ICCV）上*。

本·米尔登霍尔、斯里尼瓦桑、罗德里戈奥尔蒂斯卡扬、尼马卡德米卡兰塔里、拉维拉莫莫蒂、伦恩和阿比谢克卡尔。2019.局部光场融合：具有规范采样指南的实用观点综合。*ACM图形交易（TOG），38(4)：1-14*。

本米尔登霍尔，斯里尼瓦桑，马修坦西克，乔纳森巴伦，拉维拉莫莫蒂，和任吴。2020.Nerf：表示场景为视图合成的神经辐射场。*arXiv预印arXiv：2003.08934*。

阮福、李川、卢卡斯、理查德、杨永亮。2019.从自然图像中进行三维表示的无监督学习。*发表在IEEE国际计算机视觉会议论文集中，第7588-7597页*。

阮福、李川、巴拉班、杨永亮。2018.渲染网：用于可区分三维形状的深度卷积网络。*神经信息处理系统（NIPS）的研究进展。*

迈克尔·尼迈耶，梅切德，迈克尔·奥克斯尔和安德烈亚斯·盖格。2019.可微体积渲染：在没有三维监督的情况下学习隐式三维表示。*arXiv预印arXiv：1912.07372。*

郑俊公园、彼得·弗洛伦斯、朱利安·斯特劳布、理查德·纽科姆和史蒂文·洛夫格罗夫。2019.Deepsdf：学习连续有符号的距离函数的形状表示。*计算机视觉和模式识别国际会议（CVPR）。*

彭，迈迈尔，李。梅切德，马克·波勒菲斯和安德烈亚斯·盖格。2020.卷积占用网络。ArXiv，abs/2003.04618.

史蒂文·鲁宾和特纳说。1980.一种可快速渲染复杂场景的三维表示法。*在第七届计算机图形和交互技术年度会议论文集上，第110-116页。*

齐藤俊助、曾黄、夏敏敏、森岛重雄、金泽角、李浩。2019.Pifu：像素对齐的隐式功能，用于高分辨率覆盖的人类数字化。*2019年IEEE/CVF国际计算机视觉会议（ICCV），第2304-2314页。*

齐藤顺助、西蒙、萨拉吉和俊俊。2020.Pifuhd：多级像素对齐的隐式功能，用于高分辨率的三维人类数字化。*在IEEE关于计算机视觉和模式识别的会议的论文集上。*

沈先生和唱冰康。2000.回顾一下基于图像的渲染技术。*在视觉通信和图像处理2000年，第4067卷，第2-13页。*国际光学和光子学学会，SPIE。

文森特·西茨曼、贾斯图斯·蒂恩斯、费利克斯·海德、马提亚斯·尼斯纳、戈登·韦茨斯坦和迈克尔·佐尔霍弗。2019a.深度体素：学习持久的3d特征嵌入。*在计算机视觉和模式识别（CVPR）中。*

文森特·西茨曼、迈克尔·佐尔霍弗和戈登·韦茨斯坦。2019b.场景表示网络：连续的三维结构感知的神经场景表示。*《神经信息处理系统的进展》，第1119-1130页。*

理查德·塞利斯基。2010.*计算机视觉：算法和应用程序。*施普林格科学与商业媒体。

A.特瓦里，O。炸，J。Thies，V。西茨曼，S。伦巴第，K。Sunkavalli，R.马丁-布鲁拉，T。西蒙J. Saragih，M。Nießner，R.潘迪。Fanello，G.韦茨斯坦，JY。.-朱，C。Theobalt，M.Agrawala，E.谢克特曼，D。B高盛和M。Zollhofer。2020.神经渲染的技术现状。*计算机图形学论坛（EG STAR 2020）。*

贾斯图斯·特利斯，迈克尔·佐尔霍弗和马提亚斯·Nießner。2019.延迟的神经渲染：使用神经纹理进行图像合成。*ACM图形事务处理，38。*

阿什什·瓦斯瓦尼、诺姆·谢泽尔、尼基·帕尔马、雅各布·乌斯科里特、琼斯、艾丹·戈麦斯、Łukasz凯泽和伊利亚·波洛苏欣。2017.你所需要的就是注意力。*在神经信息处理系统的进展中，第5998-6008页。*

颜新辰、杨姬、玉美、郭一杰、李宏乐。2016.透视变压器网：在没有三维监督的情况下学习单视角三维物体重建。*《神经信息处理系统的进展》，第1696-1704页。*

姚姚、罗子欣、李世伟、张景阳、任玉凡、周磊、田方、龙泉。2020.一个用于广义多视图立体声网络的大规模数据集。*计算机视觉和模式识别（CVPR）。*

张和陈津汉。2004.基于图像的渲染、采样、压缩的调查。*信号处理：图像通信，19(1)：1-28。*

张理查德，菲利普伊索拉，阿列克谢埃弗罗斯，伊莱谢克特曼，和王奥利弗。2018.深度特征作为一种感知度量的不合理有效性。*在CVPR*。

周廷辉、塔克、弗林、菲夫和诺亚·斯纳弗利。2018.立体声放大：使用多平面图像进行学习视图合成。*在签名*。