神经辐射场（NeRFs）：回顾和一些最新发展

Mohamed Debbagh
加拿大蒙特利尔麦
吉尔大学
mohamed.debbagh@mail.mcgill.ca

***摘要*--*神经*辐射场（NeRF）是一个框架，它以全连接神经网络的权重来表示三维场景，即多层感知（MLP）。该方法是为了完成新的视图合成任务而引入的，能够从一个给定的连续视点实现最先进的逼真图像渲染。由于最近的发展扩大了基础框架的性能和能力，NeRFs已经成为一个流行的研究领域。最近的发展包括需要更少的图像来训练视图合成模型的方法，以及能够从无约束和动态场景表示中生成视图的方法。**

***Index Terms*-*volume* rendering, view synthesis, scene representation, deep learning**

I. 简介

计算机视觉研究中探索的主要问题之一是视图合成，它在计算机图形和三维渲染领域有许多影响和共享方法。这个问题的解决方案旨在开发一种方法，能够在给定来自稀疏视点集的二维RGB图像输入的情况下生成特定场景的新视图。这样一个模型的输出应该在一个连续的视点集上进行采样，从而产生同一场景的真实的新视图。一些流行的方法包括光场插值，通过基于网格的近似进行表面估计，以及最近的神经体积渲染（基于神经网络的方法）。神经辐射场（NeRF）是由Mildenhall等人提出的，属于后一类方法，它使用神经网络结构来表示场景，并使用神经体积渲染来合成新的视图，以达到最先进的效果。原文[1]将NeRF作为视图合成的突出方法加以推广，并做出了3个总体贡献，使该框架能够产生逼真的输出，能够对场景的复杂形状和表现进行建模。(1) 第一个贡献是通过一个简单的全连接神经网络来代表一个连续场景，该网络将5D输入（3个欧氏坐标维度和2个观察方向维度）映射到4D输出（RGB颜色通道和体积密度）。(2) 第二个贡献是使用神经体积渲染技术，利用可微分的相机射线，使RGB表示的优化成为可能。(3) 最后，利用位置编码技术来转换输入的

域变成一个更高的维度空间，这使得神经网络在训练期间能够捕捉到场景表征中更高频率的细节。此后，NeRF模型得到了改进和扩展，以捕捉各种模式的表示。本文回顾了原始的NeRF框架，被称为vanilla NeRF，并进一步探讨了为扩展基线模型而做出的许多贡献中的一些。这一回顾将包括以下基于NeRF的发展：PixelNeRF、RegNeRF、Mip-NeRF、Raw NeRF和NeRF in-the-Wild。为了在高层次上回顾这些概念，我们将不包括为实验而设计的具体方程或模型架构，我们建议你在原始论文中探索具体实现的细节。

II. 神经辐射场

A. 神经体积渲染

NeRF表征是建立在神经体积之上的，这是一种隐含的三维场景的体积表征，被学习并存储为一个深度神经网络的权重。在Lombardi等人[2]中，二维图像被送入一个变异自动编码器（VAE），并被编码为一个潜伏代码。解码器的输出将潜伏代码重建为一个体积体素表示，在空间的每个点上都有一个RGB和alpha通道的复合表示。虽然他们的研究目标是从二维图像中构建一个三维表示，但VAE是通过从体素表示中重建二维图像来训练的，使用光线行进技术进行体积渲染。光线行进是一个可微分的过程，它使使用梯度下降方法进行优化成为可能。场景的二维重建可以通过估计图像平面内每个像素的辐射度值来进行，这些像素是在给定的观察方向上从三维场景中投影出来的。

每个像素点的射线在垂直于图像/摄像机平面的给定观察方向上投射到三维空间，并用于表示沿射线的空间体积或占有率。像素的体积密度是通过沿射线取体积的积分来确定的；这个过程被称为体积渲染。由于在计算上不可能确定沿连续射线的体积，所以沿射线的离散点的体积被取样来估计积分；这种技术用于

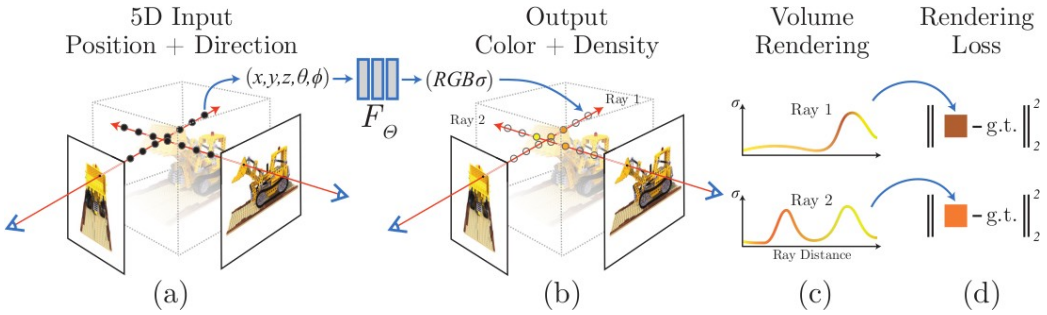

图1.NeRF场景表示管道的概述。(a) 5D输入的前馈传递。(b) 4D输出在2D空间的映射。(c) 体积渲染的光线行进。(d) 对重建损失的优化。[1]

体积渲染被称为射线行进（ray marching）。在重建的二维图像中，像素的辐射度和体积通常被表示为颜色（r,g,b）和不透明度。这个过程可以通过神经网络进行映射，这种表示方法的全部内容被称为神经体积渲染。在Lombardi等人的案例中，神经体积渲染被用来从VAE的输出中获得的三维体素输出中重建图像。然而，由于VAE的性质，当来自低维空间的潜伏代码被上采样到高维三维体素空间时，会出现伪影和重建翘曲。各种额外的技术需要被应用来减轻这些影响。因此，最终的三维形状几何往往不那么完美。相比之下，视图合成的目的是为了从新的视点生成逼真的图像。使用VAE来重建三维体素解释是没有必要的，因为这将导致如[2]中所述，不完善的地方。

B. NeRF三维场景表示法

最初的NeRF论文提出，场景的表示是一个神经量，由一个简单的全连接神经网络架构（称为多层感知器（MLP））的权重来描述，其5D输入（x、y、z、ϑ、Φ）对应于三维空间的位置、$x=(x, y, z)$和二维观察方向，$d=(\theta, \Phi)$，即对应于沿摄像机射线的一个点。的输出MLP对应于色彩通道的映射，$c=(r, g, b)$和该视角下二维图像平面上的像素的体积密度，σ。与之前的研究不同，场景的三维表示完全通过简单的前馈MLP的权重来隐含表示，而不是通过体素表示。MLP前馈网络可以表示为F_Θ：$(x, d) \rightarrow (c, \sigma)$。该MLP的参数$\Theta$是用一个可微调的参数来优化的。体积渲染函数，并在一组地面实况的基础上进行训练。图像和它们的已知观察方向。损失函数可以通过评估真实像素颜色和体积渲染过程中预期像素颜色之间的差异来选择。在论文中，作者使用了一个简单的平均平方误差。关于原始论文[1]中NeRF场景表示的视觉概述，见图1。

C. 位置编码

当按照上一节所述直接训练MLP，$F_\Theta : (x, d) \rightarrow (c, \sigma)$ 时，该模型往往难以输出高度详细的结果。这是许多编码任务中的一个常见问题，在这些任务中，人们想要编码一个通过全连接神经网络的权重来表示，如图像。这项任务是困难的，因为MLP的偏向于更快地学习低频。这意味着，这些网络往往在想要概括结果的任务上工作得更快，并避免过度拟合数据。然而，由于神经体积渲染的目标是将一个精确的几何形状适合于三维场景，所以网络最好过度拟合数据。[3] Tancik等人介绍了一种常用于变压器的方法，称为位置编码，将低频输入映射到高频域。将输入映射到高频域允许MLP捕捉到场景中的高频和高分辨率细节。当应用于MLP时，NeRF模型成为F_Φ：$(\gamma(x), \gamma(d)) \rightarrow (c, \sigma)$。其中$\gamma(.)$是一个函数，该函数将我们的输入映射到高频域。在这种情况下NeRF一个傅里叶特征映射被用来作为高频特征映射函数。请注意，这对于实现NeRF模型获得的逼真效果是必要的。

D. 财产

到目前为止，NeRF模型及其用于视图合成的优化方法已经被描述为一种神经体积再现，可以捕捉三维场景的高频几何细节。这给了NeRF一些有趣的内在属性，这些属性超出了视图合成的任务范围。要注意的第一个属性是，由于三维几何表示被存储为全连接神经网络中的权重，NeRF可以被认为是一种三维模型的压缩格式。三维模型可以通过在预先定义的视点上查询预先训练好的NeRF，然后应用三维几何构造方法（如行进立方体）来重建。这一点很重要，因为NeRF文件的大小会比模型训练的单一图像小。第二个属性是NeRF捕获了关于场景的关系性几何信息。这使我们有能力将详细的几何信息用于

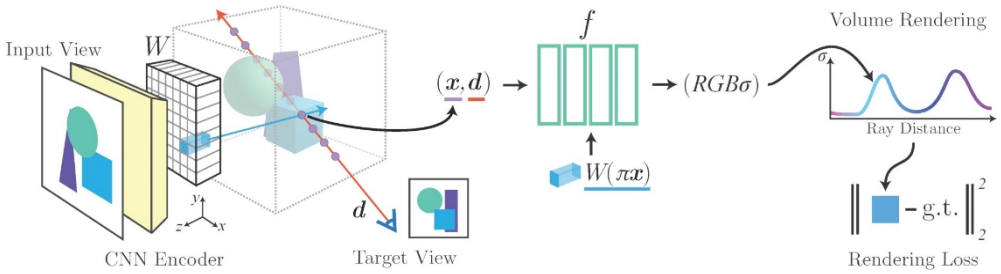

图2.PixelNeRF输入和体积渲染管道的概述。[4]

任务，如生成深度图和形状可视化。这也可以用来捕捉混合现实场景的闭塞效应。第三个属性是将不同观看方向的色彩感知效果可视化的能力。这个属性允许在给定一个固定位置的情况下，以逼真的方式捕捉各种照明条件下的场景。

E. 实施和挑战

从NeRF方法中得到的视点是非常详细的，并且在综合建模的场景和真实场景中都优于先前的先进方法。然而，到目前为止，所描述的虚无缥缈的NeRF模型在现实世界中的实施有几个限制。我们将关注的第一个方面是训练和优化过程。实现NeRF模型的挑战是，每个场景都需要在已知视点方向的图像上进行训练。虽然这个问题对于随处可见的应用来说似乎很有局限性，但是存在估计这些参数的方法，包括COLMAP结构-从运动包[5]。这可能会在新场景的生成过程中引入一些变化，尽管如此，获得的结果仍然是相当令人印象深刻的。与其他方法相比，训练和渲染过程非常缓慢，并且需要来自独特视角的多样化图像集来捕捉无缝的连续视图合成。大多数实施方案需要至少80张图像进行训练。用非常稀疏的图像训练出来的模型将产生无法解释的场景，并且无法进行概括。其他的挑战包括对所捕捉的场景的限制。由于动态因素的影响会对视图合成产生剧烈的影响，因此NeRFs被限制在静态场景中。这包括反射、移动物体和背景。香草模型所观察到的这些挑战创造了一个新的研究领域，特别是对基本的NeRF概念进行优化和扩展。我们将在本文接下来的几节中探讨解决其中一些挑战的最新发展。

III. 从更少的图像中查看合成

NeRF研究的最新发展所解决的一个挑战领域是场景的校准过程。由于训练和渲染新场景的时间和计算量很大，NeRF的实施往往受到限制。一种方法是减少用于校准过程的资源量。有两篇论文指出解决了减少所需校准图像数量的问题。

A. 像素尼弗（PixelNeRF

普通的NeRF模型需要许多来自不同视角的图像，因为MLP模型不能很好地概括。MLPs也没有纳入空间信息，因为图像在被送入训练过程之前就被压扁了。如果有一个以上的观点被用来校准场景，那么香草方法就不会考虑从多个观点中学习到的信息。当图像采样不一致且稀疏（少于80张）时，这导致了场景合成的退化。Yu等人[4]对基本的NeRF模型进行了扩展，在校准过程中加入了场景先验因素。这个模型被命名为PixelNeRF，对NeRF框架的主要贡献是通过卷积神经网络（CNN）对输入图像进行调节，以训练场景预设。为了更好地说明这一点，论文中给出了体积渲染管道的可视化概述[4]，见图2。这使得模型的训练只需要一张校准图像，尽管这只推荐用于简单的几何形状。在多视图校准中（2张或更多的图像），每个输入图像的CNN在不同视图下的输出被合并，然后再通过体积渲染过程进行反馈。PixelNerf能够在ShapeNet数据库[6]中的简单合成模型上实现连续的场景表示，只需要一张校准图像。该模型还在真实图像上进行了测试，能够使用单一的校准图像生成连贯的场景几何表征，这一点从vanilla NeRF中是无法实现的。然而，结果并不完美，并产生了假象和失真。在增加多个视图（2-3个以上）进行校准的情况下，这个问题得到了明显的缓解。

B. 摄制组

Niemeyer等人介绍了一种方法，该方法减少了浮动伪影和图像不一致的情况，这种情况在虚构的NeRF只在少数图像上训练时发生。该论文通过对未见过的视图中的斑块进行几何平滑度和颜色的正则化来实现这一目标[7]。本文所介绍的模型被命名为RegNeRF，它改进了vanilla NeRF模型的优化过程。虽然虚构的NeRF模型对重建损失进行了优化

在输入图像中，它没有被优化以学习各点的几何一致性，因此，当样本图像变得稀疏时，该方法会恶化。RegNeRF从未见过的视点的斑块中取样射线，然后定义一个优化，目的是使斑块的几何平滑度和颜色相似度正规化。这是在训练过程中通过定义颜色和几何斑块的正则化项的损失函数完成的。本文的结果显示，与以前的模型相比，在减少浮动伪影方面有明显的改进。由于RegNerf保持了原始NeRF模型的MLP架构，它在预训练中的计算成本比基于CNN的pixelNerf低。RegNeRF可以使用低至3张校准图像进行训练。

IV. 动态和非约束性条件

一个特定场景的动态条件是影响其表现的一个主要因素。通常情况下，普通的NeRF模型无法利用这些动态条件，事实上，它需要对场景进行约束，以实现没有浮动伪影和混叠等伪影的体积化渲染。NeRF模型的最新发展已经探索了利用、控制和操纵场景条件的各个方面的方法。在本节中，我们将探讨一些论文，这些论文解决了一些领域的问题，如多尺度表示的抗混叠、图像处理管道和来自无约束的样本图像的表示。

A. Mip-NeRF

多尺度表示对许多图像处理和三维渲染任务构成了挑战。从不同尺度重现三维场景或二维图像时，往往伴随着伪影，即所谓的锯齿，这往往是由混叠引起的。在NeRF模型中，当对较低分辨率的输入图像进行采样时，尤其可以观察到混叠。相同分辨率的视图的重建往往包含这些锯齿。用多尺度分辨率训练NeRF模型来缓解这个问题，往往不能带来明显的改善，特别是在试图重现高分辨率的视图时。Barron等人介绍了Mip-NeRF，这是NeRF方法的一个扩展，它使用射线锥来捕捉空间的体积，而不是一个无限小的点来控制场景的多尺度表示[8]。随着图像尺度的变化，单个像素从场景中捕获的信息量也在变化。因此，在每个像素上沿单点射线采样会在与邻近像素插值时造成失真，从而产生混叠效应。沿着一个区域的锥形射线取样点可以以非线性的方式捕获体积信息。本文通过拟合一个多变量的高斯分布，沿这些射线锥体逼近这些圆锥交点。由于取样不再沿线进行，在分布中选择样本相当于位置编码的预期值，这反过来又让网络根据调整后的空间体积从比例上进行推理。A原始论文[8]中的图表给出了锥形射线的视觉表现，见图3。这项研究的结果表明，与以前的香草NeRF方法相比，Mip-NeRF的性能优于多尺度重构。与超级采样法相比，它的计算效率也显著提高，结果相当。

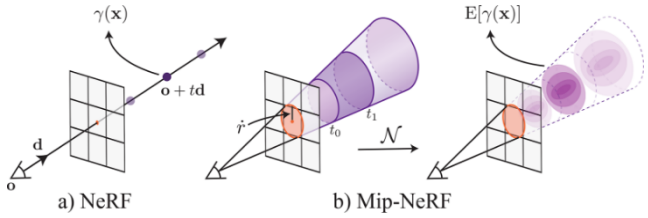

图3.NeRF射线行进和mip-NeRF射线锥高斯期望采样的比较。[8]

B. 原始NeRF

在这一节中，我们看一下NeRF模型的方法，它考虑了图像处理和后处理管道，而不是模型结构，以便从场景中获得更多的信息，从而获得令人印象深刻的结果。NeRF模型通常使用低动态范围图像（LDR）进行训练，以进行新的视图合成。这种处理程序通常是为了去除图像中的噪声，特别是在黑暗中。然而，这是以场景中较暗区域的细节损失为代价的。这种细节的损失反映在NeRF模型生成的新视图上。例如，在非常低的照明条件下的场景，会产生非常黑暗的视角图像，几乎没有细节。相比之下，高动态范围（HDR）图像利用一种技术，通过结合不同曝光或视图的多个图像来捕捉细节，甚至应用后期处理技术来重新聚焦。Mildenhall等人在他们的论文[9]中提出，NeRF模型的输入是原始的、经过最小化处理的、嘈杂的马赛克线性图像，以捕捉场景的更多细节，特别是在黑暗中。然后，NeRF可以合成场景的新的视图点，并应用后期处理技术，在最终合成的视图中捕捉类似HDR的效果。原始NeRF管道的可视化表示来自原始论文[9]，见图4。这种方法对新颖的视图合成有很多影响。首先，Raw NeRF能够生成一个去噪的场景视图，其性能优于LDR处理中使用的深度去噪方法以及多视图去噪。Raw NeRF能够在非常低的光照条件下渲染一个具有照片般真实细节的场景。此外，HDR色彩空间的后期处理方法可用于实现进一步的效果，如重构场景的曝光、色调映射和重新聚焦。在进行这些处理的同时，还能捕捉到场景的三维几何细节。

C. 野外的NeRF

香草型NeRF模型及其许多变体的一个限制是对受限采样的要求。

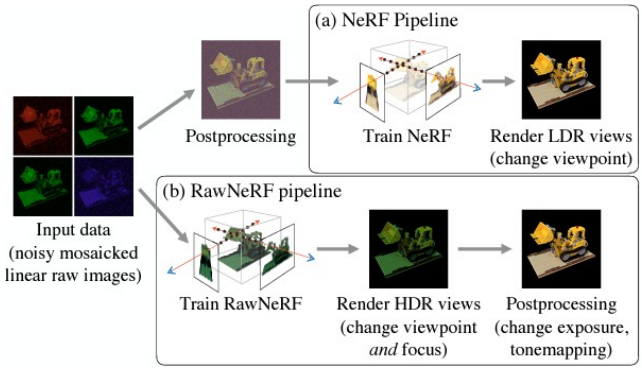

图4.原始NeRF输入和体积渲染管道的概述。[9]

条件。这限制了NeRF在真实世界和自然图像上的许多可能应用。这也限制了NeRF模型在有一个或几个物体的固定场景中的应用，并且需要相当一致的图像视点来校准。当对无约束的图像和动态场景进行训练时，NeRF会产生带有浮动假象的视图，因为它确实知道如何解释这些变化的实体。这些动态因素包括光度变化，如照明条件和天气条件，以及瞬时物体，如移动物体和临时结构。Martin-Brualla等人提出了一个被称为NeRF-W的NeRF模型的扩展，该模型嵌入了场景的静态和瞬时成分，以生成动态条件下的新视图[10]。NeRF-W能够通过对模型的输入进行外观嵌入和瞬时嵌入的调节来分解所学的静态成分和动态因素。在训练过程中，NeRF-W通过优化嵌入以及NeRF在重建损失上的权重来学习这些解释，该损失是由不确定性因素调制的。这样一来，NeRF-W就能够成功地将场景的结构与动态方面分离出来。由于瞬时嵌入是学来的，场景可以在各种条件下从训练数据的多样性中重新创造出来。从本质上讲，NeRF-W是原始NeRF模型在动态因素条件下的一个分解版本。

V. 结论

自2020年开发NeRF框架以来，已经做了许多变体和扩展，极大地提高了其性能和能力。该模型能够实现最先进的结果和照片般逼真的渲染，为这样一个框架在视图合成领域和其他领域提供了许多机会。此后，NeRF已经成为一个独立的研究领域，并不断取得重大进展。NeRFs的应用包括：电影摄影中的3D场景渲染、3D图形生成、虚拟渲染和场地漫步等等。本文涵盖了对NeRF基本框架的回顾，并探讨了到目前为止（在撰写本文时）的一些最新发展。强烈建议对每个NeRF模型的变体进行观察通过在各个项目现场的视频演示，可以直观地了解到这些能力。

参考文献

[1] 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, " in *ECCV*, 2020.

[2] S.Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y.Sheikh, "Neural volumes：从 图像中学习动态可渲染的体积，" *ACM Trans.Graph.*, vol. 38, no.4, pp. 65:1-65:14, Jul. 2019.

[3] M.Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier特征让网络在低维 域学习高频函数，" *NeurIPS*, 2020.

[4] A.Yu, V. Ye, M. Tancik, and A. Kanazawa, "pixelnerf: Neural radiance fields from one or few images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition （CVPR）*, June 2021, pp.

[5] J.L. Scho¨nberger和J.-M. Frahm，"重新审视来自运动的结构"，在计算机视觉和模式识别会议（CVPR）上，。Frahm，"重新审视来自运动的结构"，在*计算机视觉和模式识别会议（CVPR）*上， 2016。

[6] A.X. 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：一个信息丰富的三维模型库，"斯坦福大学-普林斯顿大学-丰田技术研究所在 Chicago, Tech.arXiv:1512.03012 [cs.GR], 2015.

[7] M.Niemeyer, J. T. Barron, B. Mildenhall, M. S. M. Sajjadi, A. Geiger, and N. Radwan, "Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition （CVPR）*, 2022.

[8] J.T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti aliasing neural radiance fields, " *ICCV*, 2021.

[9] B.Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, "NeRF in the dark: High dynamic range view synthesis from noisy raw images, " *CVPR*, 2022.

[10] 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 *CVPR*, 2021.