引文格式：汪骏,李赞,王和平,等. 一种高鲁棒性神经辐射场多视角重构网络[J]. 地理空间信息, 2025, 23(4):16-19 . doi:10. 3969/j.issn. 1672-4623.2025.04.004

一种高鲁棒性神经辐射场多视角重构网络

汪　骏[1]，李　赞[1]，王和平[1]，孙怀木[1]，李俊磊[1]
（1. 国网电力空间技术有限公司，北京 102209）

摘　要：神经辐射场在新视图合成领域取得了较好的效果，该技术通过训练视图与场景渲染图间的光度差学习场景中的颜色与密度场，利用足够多的视图集进行训练，可以从任意摄像机位置生成新视图。然而，现有方法受限于场景的几何与颜色特征，可能导致重建后的场景出现偏差，为此提出一种高鲁棒性神经辐射场多视角重构网络。给出一种网络结构，学习场景中的几何与颜色，以此预测颜色和深度联合概率分布梯度，用于规范场景中的几何与颜色信息，并通过引入模糊感知深度估计模块来处理单目深度分布的多模态性。在LLFF和DTU数据集上进行实验，结果表明，该方法提高了重建几何图形的质量，改善了对新视角重构的泛化性能，验证了方法的有效性。

关键词：多视角重构；神经辐射场；联合概率分布；深度估计
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2025）04-0016-04

Highly Robust Multi-view Reconstruction Network for Neural Radiation Field

WANG Jun[1], LI Zan[1], WANG Heping[1], SUN Huaimu[1], LI Junlei[1]
(1. State Grid Power Space Technology Co., Ltd., Beijing 102209, China)

Abstract: The neural radiance field has achieved good results in the field of new view synthesis. This technique learns the color and density fields in the scene by training the difference between the view and scene rendering image. By using a sufficient number of view sets for training, it can generate new views from arbitrary camera positions. However, the existing methods are limited by the geometric and color features of scene, which may lead to deviations in the reconstructed scene. Therefore, we proposed a highly robust multi-view reconstruction network for neural radiance field. This network gives a network structure that learns geometric and color features in the scene to predict the joint probability distribution gradient of color and depth, which is used to regulate geometric and color information in the scene, and introduced the module of fuzzy perception depth estimation to deal with the multi-modality of monocular depth distribution. Experiments on LLFF and DTU datasets show that this method improves the quality of reconstructed geometric graphics and the generalization performance of new views, verifying the effectiveness of method.

Key words: multi-view reconstruction, neural radiance field, joint probability distribution, depth estimation

目前，神经辐射场、神经隐含面和基于坐标的场景表征已成为视图合成和三维重建任务的研究热点。神经辐射场（NeRF）[1]通过多层感知器学习特定场景，当给定任意三维点和观察方向时，可预测密度和颜色。这种方法允许从任意视角进行可微分渲染，在渲染过程为沿射线方向预测颜色分布，并根据渲染结果进行合成，训练模型用于重建相机的摄影位姿。仅使用光度重建损失进行训练，NeRF也能获得较高的泛化能力，从而提高虚拟现实[2]、增强现实[3]和视觉特效技术的性能。然而，在输入视图数量较少的情况下，场景颜色和几何场的约束严重不足，可能生成低质量和不可信的几何图形和表面。例如，“漂浮物”就是一种常见的伪影，拟合密度场包含漂浮在半空中的半透明物质云，从训练视图渲染后，这些物质在二维视图中看起来是合理的，但在新视图中看起来却难以理解。为解决该问题，诸多学者提出多种正则化和先验学习的方法[4-11]。

本文提出一种网络结构，能够更好地学习场景中的几何与颜色，以此预测颜色和深度联合概率分布梯度，用于规范场景中的几何与颜色信息，并通过引入模糊感知深度估算模块来处理单目深度分布的多模态性，以此提高神经辐射场多视角重构质量。

1　网络总体结构

神经辐射场（NeRF）代表三维空间中的一个场，其中每个点代表具有一定不透明度的小粒子，沿不同观察方向发出不同数量的光。点的不透明度用体积密度表示，发射的光量用颜色表示。NeRF表示为2个参

第一作者简介：汪骏（1983—），硕士，高级工程师，从事直升机电力作业、电气工程工作，E-mail：Watghwn@126.com。

数化函数，即体积密度和颜色。前者将三维坐标映射到体积密度，后者将三维坐标景观方向映射到颜色域上，将颜色、体积密度和能见度的乘积沿光线进行积分，以此获得每个点的像素值。可见度用透射率来表示，透射率将体积密度指数化后的负值乘以累加，某点的透射率越高，从摄像机上看该点的可见度就越高。

参考 NeRF 网络结构，给定一组具有相机参数和姿态的场景（图 1），优化密度场 $\sigma:R^3\to R_+$ 和颜色场 $c:R^3\times S^2\to R^3_{[0,1]}$。其中密度场可在任意三维坐标 $(x,y,z)\in R^3$ 下求解，颜色场可在任意三维坐标和观察方向 $d\in S^2$ 上求值。

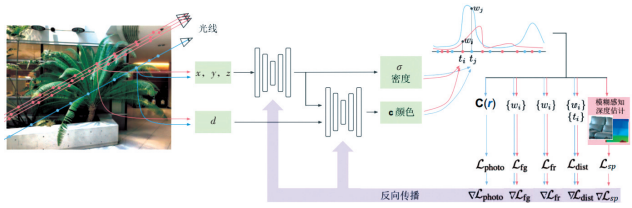


图 1　网络总体架构

在任意场景视图中，射线 $r(t)=o+td$ 的预期颜色 $C(r)$ 可通过离散样本 $t_{0:N}$（其中 $t_{i+1}>t_i>0$）获得：

$$C(r)\approx\sum_{i=1}^{N}w_i c\left(r\left(t_i\right),d\right)+\left(1-\sum_{i=1}^{N}w_i\right)c_{bg} \tag{1}$$

其中将颜色贡献的权重为 $w_i=T\left(t_i\right)\rho\left(t_i\right)$ 定义为：

$$\rho\left(t_i\right)=1-\exp(-\sigma(r(t_i))(t_{i+1}-t_i)) \tag{2}$$

$$T\left(t_i\right)=\prod_{j=1}^{i-1}\left(1-\rho\left(t_j\right)\right) \tag{3}$$

式中，exp(*) 为自然常数；e 为底的指数函数；σ 为方差。即射线 $r(t)$ 从摄像机中心 o 出发到达目标时不被吸收的概率，可计算出预期深度为：

$$D(r)=\frac{1}{2}\times\frac{\sum_{i=1}^{N}w_i t_i}{\sum_{i=1}^{N}w_i} \tag{4}$$

对密度场和颜色场进行优化，减少光度重建损失。例如，输入图像与来自相同视图渲染图像间的 L_2 差值为：

$$L_{\text{photo}}(\sigma,c)=\sum_{I}^{i=1}I_i-\frac{1}{2}wC_i \tag{5}$$

式中，C 为背景颜色，将其设置为白色。颜色贡献权重 w 正则化为紧凑分布：

$$L_{\text{dist}}=\frac{1}{D(r)}\left(\sum_{i,j}w_i w_k)(\frac{t_i+t_{i+1}}{2}-\frac{t_j+t_{j+1}}{2})^2+\frac{1}{2}\sum_{i=1}^{N}w_i^2\left(t_{i+1}-t_i\right)\right) \tag{6}$$

式中，对于靠近摄像机的几何体，正则因子的强度会增加，以权重总和为统一，这是由于在真实场景中射线被完全吸收，此时由场景的几何特征决定：

$$L_{fg}=\left|1-\sum_{i=1}^{N}w_i\right| \tag{7}$$

在只有少量视图的情况下，NeRF 经常会陷入“崩溃”，无法生成解决方案，其中每个相机都被相应训练图像全部或部分“覆盖”。为了避免该情况，本文引入一种正则化方法，即对只包含在单视图凹陷中的密度进行惩罚：

$$L_{fr}=\frac{1}{2}\sum_{i}w_i\left(n_i<=1\right) \tag{8}$$

将上述这些几何正则结合到损失函数中，如下式所示：

$$L_{\text{geom}}=L_{\text{photo}}+L_{fg}+L_{fr}+L_{\text{dist}} \tag{9}$$

2　模糊感知深度估计模块

为了充分利用深度先验，一种方法是使用深度先验来限制光线终止距离。由于 NeRF 可以表示为非透明表面，因此不存在单一光线终止距离，则通过光线终止距离分布取代光线终止距离。该累积分布函数（DF）表示光线在到达指定点之前终止的概率。事实证明，该分布的概率密度 $f_{\theta,o,d}(t)$ [12]如下：

$$f_{\theta,o,d}(t)=T_{\theta,o,d}(t)\sigma_\theta(o+td) \tag{10}$$

通过引入模糊感知深度估算模块来处理单目深度分布的多模态性，用于估计来自单一视角的深度估算的模糊性。现有单目深度估算网络预测的是单点深度，与之不同，本文将单目深度估算中固有的不确定性建模为神经网络生成的一组样本分布，并利用条件隐含最大似然法的姿态估计习多模态分布的深度估计值。图 2 给出了模糊感知深度估计模块的网络架构。具体来说，将 IMLE 与最先进的单目深度定时网络 LeReS 结合，令模糊感知深度估计模块的输入图像潜在编码为 $z\sim N(0,I)$，并为输入图像 I 输出条件深度样本 $G(I,z)$。为了向网络加入随机性，在网络骨干中加入 Ada 层，用于预测中间网络特征的规模和特征，为此在深度估计主干网的编码器中添加了 4 个 Ada 层。

为了更好地利用空间信息，本文提出一种空间损失 L_{sp}，利用 2 种多模态分布的样本来限制 NeRF 优化，目标是在每个视图的单目深度分布中找到全局一致的模式子集。通过这种方式，能够将来自不同视图的信息融合在一起，采用自多视图信息才能更好地解决固有的模糊性问题，因此可以找到在所有视图中一

致的几何形状，并将物体表面“拼接”在一起。

本文给出一种具有模式搜索的网络损失，即用于训练分布的模式为提供监督分布模式的子集交叉熵，它是从NeRF射线终止距离分布到模糊深度估计分布的交叉熵[7]：

$$H\left(f_{\theta,o,d},G_{o,d}\right)=-E_{f_{\theta,o,d}}\left[\log G_{o,d}\right] \tag{11}$$

正如IMLE论文[7]给出的结论，这种做法等同于对2个分布样本间的$L2$准则最小值进行惩罚，因此将空间信息损失记为：

$$L_{sp}(r)=\sum_{i\in[N]}\min_{j\in[M]}x_i-\left|y_j\right| \tag{12}$$

如果损耗只作用于矩，即沿光线的积分，那么沿光线的所有点只能得到相同的监督目标。因此，空间信息损失允许三维监督，而不像传统方法只能进行二维监督。三维监督允许清除空间中的噪声，由于要将光线终止距离分布中的样本向更深处移动，损失须降低其前面所有点的概率密度。优化和训练NeRF模型的总损失L（预测图像和真值图像的MSE损失）为：

$$L=L_{\text{geom}}+L_{sp} \tag{13}$$

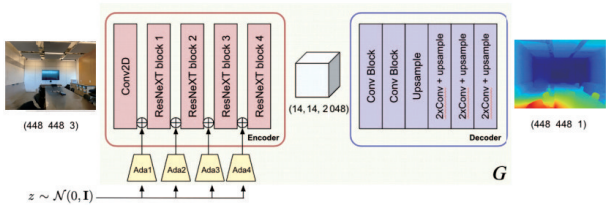


图2 模糊感知深度估计模块结构

3 实验结果与分析

实验参数中，令扩散时间参数平滑地从0.1改变到0，其他变量也相应得出。通过上述方式，当网络训练并生成越来越精确的颜色和深度时，预期输入的噪声将逐渐减小。根据经验发现，这种正则化权重的安排能产生最佳效果。在单个Nvidia 3090 GPU上，本文提出的网络在每个场景上的训练时间约为30 min。此外，有25%的时间采用训练姿态进行补丁渲染，并直接从训练图像中采样RGBD补丁的RGB分量，这在NeRF渲染还不够精确的早期阶段非常重要。

3.1 数据集

在LLFF和DTU两个数据集上进行了实验。LLFF[13]数据集有8个场景，每个场景有20~62张由手持摄像机拍摄的图像。少量图像用于训练，测试图像用于评估新视图合成质量。选择LLFF数据集进行评估，原因在于它可以与其他SOTA NeRF模型进行比较。DTU[14]数据集包含黑色背景下放置在桌子上的物体图像，图像和深度图由安装在工业机械臂上的结构光扫描仪采集，该数据集提供了用于评估的图像、姿势和地面实况点云。

3.2 实验分析

采用PSNR、SSIM[15]和LPIPS等[16]图像相似度指标评估结果，并计算每个场景的平均指标值：

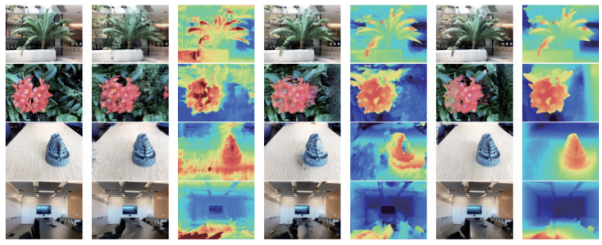
$$\text{Average}=\sqrt[3]{10^{-\text{PSNR}/10}\times\sqrt{1-\text{SSIM}}\times\text{LPIPS}} \tag{14}$$

在几何估计任务中，采用立方体将密度场的等距面转换为网格。对网格进行剔除，只保留在至少一个训练视图中可见的部分，屏蔽背景表面。对网格采样生成点云，并估计点云与地面实况点云之间的平均倒角$L1$距离。表1给出了在LLFF和DTU数据集上，使用3、6和9个视图进行训练的结果。本文方法的几何基线模型与SOTA方法相比，当视图数较少时，正则化器会对最终结果产生较大影响，因此可以更容易地比较正则化器。由表1可见，几何基线与本文方法均优于其他方法，在PSNR、LPIPS和Average指标上都

表1 本文方法与SOTA在LLFF和DTU数据集上的视图合成结果对比

数据集	方法	设置	PSNR ↑			SSIM ↑			LPIPS ↓			Average ↓		
			3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
LLFF	mip-NeRF	Optimized per Scene	14.62	20.87	24.26	0.351	0.692	0.805	0.495	0.255	0.172	0.246	0.114	0.073
	DietNeRF	Optimized per Scene	14.94	21.75	24.28	0.370	0.717	0.801	0.496	0.248	01.83	0.240	0.105	0.073
	PixelNeRFft	DTU + ft per Scene	16.17	17.03	18.92	0.438	0.473	0.535	0.512	0.477	0.430	0.217	0.196	0.163
	MVSNeRFft	DTU + ft per Scene	17.88	19.99	20.47	0.584	0.660	0.695	0.327	0.264	0.244	0.157	0.122	0.111
	RegNeRF	Optimized per Scene	19.08	21.10	24.86	0.587	0.760	0.820	0.336	0.206	0.161	0.146	0.086	0.067
	Geometric Baseline	Optimized per Scene	19.88	24.28	25.10	0.590	0.765	0.802	0.192	0.101	0.084	0.118	0.071	0.060
	(本文方法)	Optimized per Scene	19.98	24.34	25.04	0.578	0.777	0.765	0.219	0.103	0.086	0.137	0.086	0.073
DTU	mip-NeRF	Optimized per Scene	8.68	16.54	23.58	0.571	0.741	0.879	0.353	0.198	0.092	0.323	0.148	0.056
	DietNeRF	Optimized per Scene	11.85	20.63	23.83	0.633	0.778	0.823	0.314	0.201	0.173	0.243	0.101	0.068
	PixelNeRFft	DTU + ft per Scene	18.95	20.56	21.83	0.710	0.753	0.781	0.269	0.223	0.203	0.125	0.104	0.090
	MVSNeRFft	DTU + ft per Scene	18.54	20.49	22.22	0.769	0.822	0.853	0.197	0.155	0.135	0.113	0.089	0.069
	RegNeRF	Optimized per Scene	18.89	22.20	24.93	0.745	0.841	0.884	0.190	0.117	0.089	0.112	0.071	0.047
	Geometric Baseline	Optimized per Scene	13.60	16.43	22.01	0.661	0.759	0.853	0.212	0.147	0.071	0.185	0.092	0.056
	(本文方法)	Optimized per Scene	18.99	20.24	25.18	0.699	0.850	0.884	0.161	0.094	0.045	0.136	0.051	0.034

取得最佳值。几何基线法在LLFF方面的指标较高，与几何基线方法相比，本文方法生成的深度更为准确。测试视图中包含的部分场景在任何训练视图中均不可见，这些被遮挡的场景部分会严重影响重建得分，表2给出了对重建质量的评估。当NeRF与所有视图拟合时，在DTU数据集的15个扫描结果中，当有大量视图的情况下，由于训练视图能提供更多的场景信息，先验值的作用并不明显。尽管如此，先验信息不会导致伪影现象，能够帮助处理无纹理表格等模糊区域。实验尽管是在室内房间大小的场景图像上进行训练，但在以物体为中心的重构任务中本文方法更具鲁棒性。



a 真值　b RegNeRF　c Geometric Baseline　d 本文方法

图3　在LLFF数据集上视图合成任务的定性结果比较

表3给出了在LLFF和DTU数据集上的评估结果，每个优化项对新视图合成和重建质量如表3所示，几何基线法在LLFF数据集上得分较高，但在DTU数据集中得分不高。图3中的定性结果表明，即使视觉得分较高，几何基线估计法获得的几何形状仍具有一定的偏差。本文方法在DTU上的得分较高，但在新视图合成指标上较低，这是由于该方法倾向于在场景中引入细节。图4给出一组复杂纹理场景下的实验，由实验结果可见，与其他方法相比，本文方法能够更好地获得复杂纹理场景下的新视角生成结果。表3中给出的消融实验结果验证了本文方法中的每个损失项的有效性。



a 真值　b RegNeRF　c Geometric Baseline　d 本文方法

图4　复杂纹理影像定性结果比较

表2　在DTU数据集的几何重建结果对比

基于SDF的方法	*L*1均值	基于NeRF的方法	*L*1均值
UNISURF	1.02	Instant NGP	1.71
NeuS	0.84	NeRF	1.49
VolSDF	0.86	Geometric Baseline	1.36
MonoSDF	0.73	本文方法	1.68

表3　消融实验

方法	LLFF			DTU			
	3视角	6视角	9视角	3视角	6视角	9视角	所有视角
$\nabla L=\nabla L_{\text{photo}}$	0.220	0.124	0.089	0.212	0.144	0.118	2.88
$\nabla L=\nabla L_{\text{photo}}+\nabla L_{fg}$	0.218	0.123	0.089	0.196	0.124	0.091	1.74
$\nabla L=\nabla L_{\text{photo}}+\nabla L_{fg}+\nabla L_{fr}$	0.146	0.083	0.071	0.216	0.124	0.092	1.73
$\nabla L=\nabla L_{\text{photo}}+\nabla L_{fg}+\nabla L_{fr}$	0.132	0.081	0.063	0.192	0.096	0.069	1.69
$\nabla L=\nabla L_{\text{photo}}+\nabla L_{fg}+\nabla L_{fr}+\nabla L_{\text{dist}}$	0.122	0.069	0.060	0.183	0.091	0.055	1.34

4　结　语

本文提出一种高鲁棒性神经辐射场多视角重构网络。给出一种网络结构，通过学习场景中的几何与颜色特征预测颜色和深度联合概率分布梯度，引入模糊感知深度估算模块来处理单目深度分布的多模态性。通过实验证明了所提出的方法可以有效提高视图合成性能，未来工作可通过采用自监督的单目深度估计或自监督立体匹配等方法进行优化。

参考文献

[1] Ben M, Pratul P, Srinivasan M T. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[J]. Communications of the ACM, 2021, 65(1):99-106

[2] 吉宸佳, 曹爽, 张杰, 等. 无人机航测虚拟仿真实验系统开发与设计[J]. 地理空间信息, 2022, 20(9):58-62

[3] 王晓庆, 丁鹏辉, 刘祥明, 等. 地理国情AR"纸智图"的研究与实现[J]. 地理空间信息, 2021, 19(6):13-16

[4] Amos G, Lior Y, Niv H, et al. Implicit Geometric Regularization for Learning Shapes[J]. CML, 2020(10):3 789-3 799

[5] Peter H, Pratul P, Srinivasan B M, et al. Baking Neural Radiance Fields for Real-time View Synthesis[J]. ICCV, 2021(8):5 855-5 864

[6] Anpei C, Ze X X, Fu Q Z, et al. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-view Stereo[J]. ICCV, 2021(6):14 104-14 113

[7] Thomas M, Alex E, Christoph S, et al. Instant Neural Graphics Primitives with a Multiresolution Hash Encod-

（下转第99页）