文物建筑三维重建新方法—FS-NeRF

刘家旺[1] 谢晓尧[1] 刘 嵩[1]
LIU Jiawang XIE Xiaoyao LIU Song

摘 要 针对文物建筑保护面临的三维重建挑战，文章研究开发了一种名为 FS-NeRF 的新模型。基于 F2-NeRF 模型，结合 S3IM 技术，通过神经辐射场（NeRF）实现高效的三维重建。利用手机影像数据，采用运动恢复结构（SfM）方法进行预处理，并引入透视扭曲和自适应空间划分技术，提高了模型的重建速度和质量。在对云南石钟山石窟数据集的实验中，FS-NeRF 在训练效率、PSNR、SSIM 和 LPIPS 等评估指标上表现出色，并在 LLFF 数据集中展示了其良好的泛化能力。研究表明，方法不仅提升了文物建筑重建的效率和质量，还为数字化保护和虚拟展示提供了新的解决方案。

关键词 文物保护；计算机视觉；三维重建；COLMAP；神经辐射场

doi：10.3969/j.issn.1672-9528.2024.11.049

0 引言

文物古迹是人类文化遗产的重要组成部分，承载着丰富的历史、文化和艺术价值。然而，许多文物由于时间、自然因素或人为破坏，逐渐丧失了原有的历史风貌，特别是在古迹建筑中尤为常见。作为地标性建筑，其无法移动，这给文物建筑的保护带来了挑战，因此，对这些建筑进行精确而全面的重建和保护已迫在眉睫。

随着计算机技术和数字化技术的发展，文物建筑三维重建工作也进入了数字化阶段，数字化建模不仅可以还原文物建筑的外观，还能为修复和保护工作提供重要的辅助。数字化的三维模型可以用于虚拟修复实验，尝试不同的修复方案，评估其对文物建筑的影响，从而选择最合适的修复策略。这种模拟实验可以减少对实际文物的干预和风险，提高修复的效率和准确性。同时也为文物建筑的数字展示和虚拟游览提供了可能。人们可以通过互联网或虚拟现实设备，远程体验和探索这些文化遗产，了解它们的历史、文化背景和建筑风格。这种数字化展示不仅为受众提供了更丰富的体验，也为文物建筑的保护、传承和教育提供了新的途径，因此将数字化建模应用到文物重建中具有重要意义。

目前，文物重建任务中常用的还是传统的基于视觉重建方法，这种方法通常需要大量的视角图像或激光扫描数据，并且负担极高的时间成本，对于大量文物古迹的重建与更新是不够高效的[1]。而神经辐射场（neural radiance fields，NeRF）[2]技术利用神经网络模型，能够从有限的二维图像数据中重建出高精度、逼真的三维场景。对于文物建筑，这意味着可以通过少量的图像，还原出精细的结构、质地和表面细节，尤其是对于那些无法随意移动或受到空间限制的文物建筑。

本文采用手机拍摄获取历史文物建筑的数据影像，并在 F^2-NeRF[3] 模型的基础上引入 S3IM[4] 方法（即 FS-NeRF）应用于文物建筑场景中。与传统的 NeRF 重建方法相比，FS-NeRF 在重建质量、速度和资源消耗方面表现得更加出色。

1 方法

本文旨在通过基于 NeRF 的方法来对文物建筑进行三维重建，在众多 NeRF 方法中，F^2-NeRF 是一种基于网格的快速 NeRF 训练方法，能够处理任意相机轨迹。因此本文选用 F^2-NeRF 作为文物建筑三维重建的基本模型，并在此基础上引入 S3IM 方法对该模型进行改善，以达到更好的渲染效果，解决传统三维重建技术训练时间长，精度不够高的问题。具体工作流程如图 1 所示。首先，使用手机对文物建筑进行视频数据采集，再通过 colmap[5] 进行预处理获得数据集，其中主要用到文献 [6] 和文献 [7] 中所提到的运动恢复结构（SfM）方法。其次，对于给定的场景区域，根据输入的视锥体对空间进行细分。最后，对于每个子区域，构建基于可见相机的透视变形函数。密度和颜色是从同一哈希表但使用不同的哈希函数获得的场景特征向量中解码的。

1. 贵州师范大学贵州省信息与计算科学重点实验室
贵州贵阳 550001

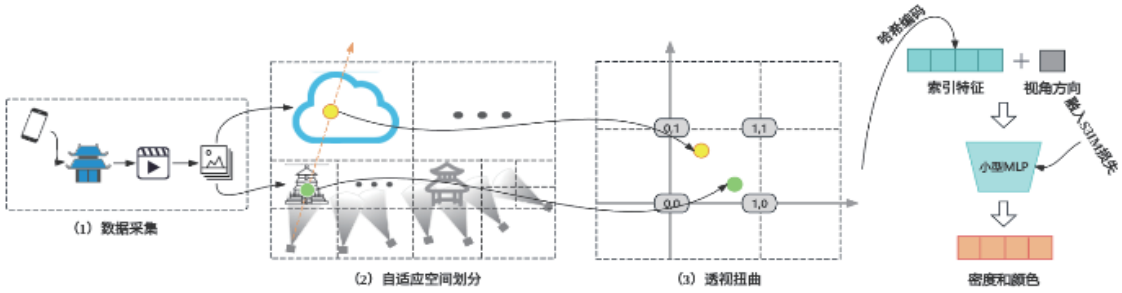

图 1 工作流程图

1.1 F^2-NeRF 算法

其核心方法是透视扭曲（perspective warping）和自适应空间划分（adaptive space subdivision）。

1.1.1 自适应空间划分

将场景空间细分，保证每个子空间的可见相机相同，以便构造相应的透视扭曲函数，并采用八叉树（octree）数据结构存储划分后的区域。

（1）以所有相机为中心，定义一个极大的初始边界框作为根节点，然后递归地进行检查和划分：

在边长为 s 的树节点上，检索其视锥体与该节点相交的所有可见相机。然后，如果存在任意一个可见相机中心，其到节点中心的距离 $d \leq \lambda s$，其中 λ 预设为 3，则该节点被细分为 8 个边长为 $s/2$ 的子节点。否则，当前节点足够小，停止细分它并将其标记为叶节点。

（2）对于每个划分的区域，选择可见的相机集合，那些超过 $n_c = 4$ 个摄像机可见的区域，通过使所选摄像机的最小成对距离尽可能大来进一步选择 n_c 个可见摄像机，并构建相应的透视扭曲函数。

1.1.2 透视扭曲

透视扭曲是该模型自定义的一种新的空间扭曲方法，能够在任意相机轨迹下处理无限场景。

（1）定义透视扭曲函数

定义透视扭曲函数 $F(x)$ 将原始的三维空间点映射到扭曲空间，其公式为：

$$F(x) = \boldsymbol{M}G(x) \tag{1}$$

式中：$G(x)$ 是将三维点 x 投影到所有相机上的二维坐标，$\boldsymbol{M}$ 是通过对二维坐标进行主成分分析[8]（PCA）得到的投影矩阵。

（2）主成分分析（PCA）

对于给定的区域 S 和相机集合 C_i，通过 PCA 分析得到投影矩阵 $\boldsymbol{M}$:

$$\boldsymbol{M} = \mathrm{PCA}([C_1(x), C_2(x), ..., C_{nc}(x)]) \tag{2}$$

式中：$[C_1(x), C_2(x), ..., C_{nc}(x)]$是点 x 在所有可见相机上的投影坐标。

1.1.3 构建网络

在扭曲空间中构建基于网格的场景表示。

（1）多哈希函数

对于每个叶子节点，使用不同的哈希函数计算网格顶点的哈希值，以减少冲突：

$$\mathrm{Hash}_i(v) = \left(\bigoplus_{k=1}^{3} v_k \pi_{i,k} + \Delta_{i,k} \right) \bmod L \tag{3}$$

式中：$\pi_{i,k}$ 和 $\Delta_{i,k}$ 是随机大素数，v_k 是网格顶点的坐标，L 是哈希表的长度。

（2）特征插值

对于给定点 x 在扭曲空间中的坐标 z，通过三线性插值从相邻的八个网格顶点获取特征向量，然后输入到一个小型 MLP 网络中计算颜色和密度。

1.1.4 采样和损失函数的设计

（1）采样点计算

通过对扭曲空间中的点进行均匀采样，可以在原始欧氏空间中获得非均匀采样，在图像上实现近似均匀采样，从而提高采样效率并带来更稳定的收敛效果。

$$\begin{aligned} x_i &= o + t_i d \\ x_{i+1} &= x_i + \frac{l}{\|J_i d\|_2} d \end{aligned} \tag{4}$$

式中：o、d 分别是该相机的原点和方向，J_i 是透视扭曲函数在采样点 x_i 处的雅可比矩阵，l 是预设的采样间隔。

（2）损失函数

$$\mathcal{L} = \mathcal{L}_{\mathrm{recon}\,(c(r),c_{\mathrm{gt}})} + \lambda_{\mathrm{Disp}}\mathcal{L}_{\mathrm{Disp}} + \lambda_{\mathrm{TV}}\mathcal{L}_{\mathrm{TV}} \tag{5}$$

$\mathcal{L}_{\mathrm{recon}\,(c(r),c_{\mathrm{gt}})} = \sqrt{(c(r) - c_{\mathrm{gt}})^2 + \epsilon}$是一个颜色重建损失函数[9]，其中$\epsilon = 10^{-4}$。$\mathcal{L}_{\mathrm{Disp}}$和$\mathcal{L}_{\mathrm{TV}}$是两个正则化损失函数。第一个是视差损失$\mathcal{L}_{\mathrm{Disp}}$，用于限制视差（逆深度）不至于过大，有助于减少浮动伪影；第二个是全变分损失[10] $\mathcal{L}_{\mathrm{TV}}$，用于使相邻八叉树节点 i 和 j 边界上的点具有相似的密度和颜色。

1.2 引入 S3IM 方法

F^2-NeRF 方法主要依赖于 MSE 损失，仅考虑了局部像

素的点级信息，未能利用远处像素的集体监督，因此无法有效捕捉图像中的结构信息，导致在图像质量感知上表现较差。为了解决这一问题，本文在 F²-NeRF 模型的基础上引入了 S3IM（stochastic structural SIMilarity）方法来改善模型性能。

S3IM 基于 SSIM 指数，该指数能够更好地反映人类视觉系统对图像质量的感知。与仅考虑像素亮度的传统方法不同，SSIM 还包含了对比度和结构信息，从而能够更全面地评估图像质量。S3IM 通过随机采样小批量像素，并计算这些像素块的 SSIM，以捕捉非局部的结构信息。利用多路复用损失，将一组像素的结构信息纳入损失函数中，从而利用图像中远距离像素之间的关系进行监督学习，有效地利用了图像中远距离像素之间的关系，提高了模型的泛化能力和鲁棒性。

1.2.1 S3IM 指数及损失

（1）S3IM 指数

设计了随机结构相似性（S3IM）指数，用于评估两个像素组的相似性：

$$\mathrm{S3IM}\left(\hat{R},R\right)=\frac{1}{M}\sum_{m=1}^{M}\ \mathrm{SSIM}\left(P^{(m)}\left(\hat{C}\right),P^{(m)}(C)\right) \quad (6)$$

式中：M 是调整 S3IM 重要性的超参数，SSIM 表示结构相似性指数，$P^{(m)}$ 表示第 m 次随机生成的像素块。

（2）多路复用损失

$$L_M(\Theta)=\mathcal{L}+\lambda L_{\mathrm{S3IM}}(\Theta,R) \quad (7)$$

式中：$\mathcal{L}$表示 F²-NeRF 方法的损失函数，λ 是调整 S3IM 重要性的超参数，$L_{\mathrm{S3IM}}(\Theta,R)$表示计算一组非局部像素的结构相似性。

1.2.2 训练流程

（1）数据采样

使用 F²-NeRF 方法从数据集中采样一个小批量的射线 R。

获取真实像素值$C=\{C(r)|r\in R\}$

计算渲染像素值$\hat{C}=\{\hat{C}(r)|r\in R\}$

（2）计算 S3IM 损失

随机生成多个像素块$P^{(m)}(\hat{C})$和对应的真实像素块$P^{(m)}(C)$。

计算每个像素块的 SSIM 值，并求平均得到 S3IM 损失：

$$L_{\mathrm{S3IM}}(\Theta)=1-\frac{1}{M}\sum_{m=1}^{M}\ \mathrm{SSIM}\left(P^{(m)}\left(\hat{C}\right),P^{(m)}(C)\right) \quad (8)$$

（3）总损失计算

结合 F²-NeRF 损失和 S3IM 损失，计算总损失：

$$L_M(\Theta)=\mathcal{L}+\lambda L_{\mathrm{S3IM}}(\Theta,R) \quad (9)$$

（4）模型更新

计算总损失的梯度$\nabla L_M(\Theta)$，并使用梯度下降 SGD 算法更新模型参数Θ。

2 实验

在文物建筑的三维重建任务中，图像数据对重建结果十分重要。文物建筑的摄影测量技术通常依赖无人机或激光扫描仪进行数据采集。然而，由于文物建筑的不可移动性，无人机 [11] 和激光扫描仪 [12] 往往无法获取某些角落的影像数据。因此，使用精细度较低的云台或手机等便携设备来采集相关数据，并重建出高质量的效果，是未来研究和发展的重要方向。

本文选取了使用手机摄影测量数据拍摄的云南省剑川县石钟山石窟作为研究对象。石钟山石窟始建于南诏时期，是研究中国佛教石窟艺术的重要遗址，作为云南地方历史和文化的宝贵资料，石钟山石窟的数字化重建具有十分重大的理论价值与文化价值。此外，本研究还为三维重建领域提供了一个具有挑战性的相机运动轨迹随机的小型文物建筑数据集，包含十五个场景。

本文分别使用 Instant-NGP[13] 和 Mip-NeRF[14] 与本文方法进行石钟山石窟的数字化重建。

2.1 实验准备

本文利用成本低、灵活性强的手机设备对石钟山石窟内不同石窟进行拍摄，从不同视角采集了石窟的影像数据。

本研究的所有实验均在相同的环境配置下完成，实验配置包括 AMD 2700X 处理器、8 GB RTX 3070 Ti 显卡和 64 GB 内存。对于数据集的制作，首先对采集到的视频数据使用基于运动分析的关键帧提取算法提取关键帧；接着对提取出的图像数据进行预处理，包括估计相机的内外参数，这一步骤是通过 COLMAP 软件来完成的。

2.2 实验结果比较

2.2.1 重建效果比较

本文采用了 NeRF 领域中较为经典的 Instant-NGP 和 Mip-NeRF 模型，以及本文提出的方法，对文物建筑数据集中的 1 号窟场景进行数字化重建比较，其对比结果如图 2 所示。图 2 左侧的前三列展示了各模型在某一视角下的重建结果，最右侧的 Ground Truth 为原始图像。可以看到，在仅迭代 20 000 的情况下，Mip-NeRF 模型对于该类运动轨迹随机的数据集表现较差，重建结果较为模糊。相比之下，Instant-NGP 模型的表现明显更好，但与本文方法相比，其重建结果仍存在一些不足，如图像光亮过曝，以及重建物体前方有少许伪影遮挡（正方框所圈部分）。总体而言，从肉眼观察来看，本文模型在颜色和形状的各个方面更接近于 Ground Truth。

（a）Mip-NeRF（b）Instant-NGP（c）FS-NeRF（d）Ground Truth

图 2 各模型重建效果对比

2.2.2 评估指标比较

对于文物建筑数据集，遵循常用的设置，将每八张图像中的一张设置为测试图像，一张设置为训练集。本文使用 PSNR、SSIM 和 $LPIPS_{VGG}$ 三个指标对文物建筑数据集中的 1 号窟场景进行模型评估，获得的结果如表 1 所示。

表 1 文物建筑数据集上的结果

方法	时间 /min	PSNR	SSIM	LPIPS
Mip-NeRF	30	22.114 4	0.549 5	0.631 6
Instant-NGP	9	**23.064 3**	0.703 5	0.332 5
FS-NeRF	**7**	21.257 8	**0.817 7**	**0.308 1**

由表 1 可知，在训练时间方面，本文的 FS-NeRF 模型表现出色，仅需 7 min 即可完成训练，是三个模型中最快的，表明其在计算效率上具有显著优势。其次，FS-NeRF 在结构相似性指数（SSIM）和感知相似性（LPIPS）两个指标上均表现突出，其 SSIM 值为 0.817 7，为三个模型中最高，显示其在捕捉图像细节和结构方面的优越性。同时，FS-NeRF 的 LPIPS 值为 0.308 1，为三个模型中最低，说明其在感知质量方面胜过其他模型。但在处理图像噪声方面，其表现不如 Instant-NGP 模型（PSNR 为 23.064 3）。

综合来看，FS-NeRF 模型在训练效率和感知质量方面具有显著优势，特别适用于对训练速度和感知质量要求较高的应用场景。然而，对于那些对 PSNR 有较高要求的应用，FS-NeRF 可能需要进一步优化，以提升其在信噪比方面的表现。

2.2.3 模型泛化性展示

为了评估所提方法的泛化能力，本文在公认的 LLFF 数据集上进行了进一步测试。测试过程中沿用了文物建筑数据集的实验参数，所得效果及评估指标如图 3 和表 2 所示。可以看出，FS-NeRF 模型的重建效果仍然是三个模型中最好的，并且在各评估指标上均优于其他模型，表明该模型具有良好的泛化能力。

（a）Mip-NeRF（b）Instant-NGP（c）FS-NeRF（d）Ground Truth

图 3 各模型重建效果对比

表 2 LLFF 数据集上的结果

方法	时间 /min	PSNR	SSIM	LPIPS
Mip-NeRF	30	22.748 7	0.614 1	0.507 7
Instant-NGP	8	26.875 6	0.803 7	0.221 4
FS-NeRF	**7**	**29.855 3**	**0.898 5**	**0.123 6**

2.3 消融实验

对文物建筑数据集中的 1 号窟场景及 LLFF 数据集中的 flower 场景进行消融研究。在消融研究中，本文使用多分辨率哈希网格作为场景表示，并改变损失函数，包括引入 S3IM 损失的 FS-NeRF 模型和未改动损失函数原有的 F^2-NeRF 模型。实验结果如表 3 所示。

表 3 消融实验结果

数据集	方法	PSNR	SSIM	LPIPS
文物建筑	F^2-NeRF	20.254 7	0.687 7	0.416 0
	FS-NeRF	**20.787 9**	**0.692 9**	**0.405 6**
LLFF	F^2-NeRF	29.775 6	0.897 7	0.124 2
	FS-NeRF	**29.855 3**	**0.898 5**	**0.123 6**

这证明了 FS-NeRF 模型在不同数据集上的多个评估指标上均优于 F^2-NeRF 模型，显示了其在图像重建方面的优越性和泛化能力。

2.4 实验结果展示

图 4 展示了 FS-NeRF 的重建实验结果，分别对应于石钟山石窟内三个不同场景下的不同视角重建效果，验证了 FS-NeRF 在各类文物建筑中的广泛适用性。

（a）Ground Truth（b）FS-NeRF（c）octree 深度图（d）深度图

图 4 更多的 FS-NeRF 重建结果

3 结语

长期以来，文物保护工作一直在稳步推进，但对不可移动的古迹和地标建筑的保护始终面临挑战。随着科技的进步，三维重建技术为这一难题提供了新的解决方案，通过数字化重建古建筑，可以更好地保护和恢复其历史面貌，并将它们转化为可永久保存的虚拟资产。本文探讨了使用 NeRF 技术进行文物建筑的数字化重建，并对优秀的 F^2-NeRF 模型进行改进，通过与其他经典的 NeRF 模型进行重建对比，本文提出的 FS-NeRF 模型在重建效果上表现更为优秀。但仍存在一定的局限性，首先，该模型在处理图像噪声方面，表现不佳；其次，如图 3 的左上角所示，对于高光有反射的部分渲染模糊。为此，需要在今后的工作中进一步完善。

参考文献：

[1] 程斌 , 杨勇 , 徐崇斌 , 等 . 基于 NeRF 的文物建筑数字化重建 [J]. 航天返回与遥感 ,2023,44(1):40-49.

[2] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[3] WANG P, LIU Y, CHEN Z X, et al. F^2-NeRF: fast neural radiance field training with free camera trajectories[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023:4150-4159.

[4] XIE Z K, YANG X D, YANG Y J, et al. S3IM: stochastic structural SIMilarity and its unreasonable effectiveness for neural fields[C]// 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023:17978-17988.

[5] SCHÖNBERGER J L, FRAHM J M. Structure-from-motio revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016:4104-4113.

[6] WU C C. Towards Linear-time Incremental Structure from Motion[C]// 2013 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2013: 127-134.

[7] CUI H N, SHEN S H, GAO W, et al. Efficient Large-scale Structure from Motion by Fusing Auxiliary Imaging Information[J]. IEEE transactions on image processing, 2015, 24(11): 3561-3573.

[8] WOLD S, ESBENSEN K, GELADI P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.

[9] BARRON J T , MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: unbounded anti-aliased neural radiance fields[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022:5460-5469.

[10] RUDIN L I, OSHER S. Total variation based image restoration with free local constraints[C]//Proceedings of 1st International Conference on Image Processing. Piscataway: IEEE, 1994:31-35.

[11] 李兵 , 岳京宪 , 李和军 . 无人机摄影测量技术的探索与应用研究 [J]. 北京测绘 , 2008(1): 1-3.

[12] BOEHLER W, VICENT M B, MARBS A. Investigating laser scanner accuracy[J]. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2003, 34(5): 696-701.

[13] MÜLLER T, EVANS ALEX, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics (ToG) , 2022,41(4): 1-15.

[14] BARRON J T, MILDENHALL B, TANCIKBARRON M, et al. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021:5835-5844.

【作者简介】

刘家旺（2000—），男，湖南邵阳人，硕士研究生，研究方向为：计算机视觉、三维建筑重建，email: ljw2569173403@163.com。

谢晓尧（1952—），男，贵州贵阳人，博士，教授，研究方向：网络通信、信息安全与人工智能，email: xyx@gznu.edu.cn。

刘嵩（1983—），男，贵州贵阳人，博士，副教授，研究方向：计算机视觉、三维重建，email:songliu@gznu.edu.cn。

（收稿日期：2024-08-13）