

# 一种基于航拍数据的空间自适应划分与 注意力引导城市自动三维重建方法

梁书凝, 谭诗瀚, 戈文一, 钟娟, 王录涛

(成都信息工程大学 计算机学院, 四川 成都 610225)

**摘要:** 基于城市神经辐射场(NeRF)场景重建的主流方法是将整个场景分割为多个区域,但在无人机航拍场景中,一旦飞行轨迹变化导致场景分布不均时,现有方法的分割效率低下,难以处理同一场景不同高度拍摄的图像细节,导致渲染质量不佳。为此,提出一种端到端的注意力引导框架。首先,利用相机位姿进行自适应空间划分,解决不均匀场景分割的效率问题;其次,从不同高度图像中提取、融合目标视图最相关的特征来增强细节,解决航拍场景高保真渲染问题。实验表明,所提模型在城市市场景数据集56Leonardv、SCUTic上均取得了最佳性能,执行渲染的速度相较于目前最高效的方法提升了65%以上。

**关键词:** 自适应空间划分;特征融合;注意力引导;NeRF模型;城市市场景

**DOI:** 10.11907/rjtk.241892

**中图分类号:** TP391.4

**文献标识码:** A

**文章编号:** 1672-7800(2026)001-0127-08



扫描二维码阅读全文:

## An Automatic Three-Dimensional City Reconstruction Method with Space Adaptive Division and Attention Guidance Based on Aerial Photography Data

LIANG Shuning, TAN Shihan, GE Wenyi, ZHONG Juan, WANG Lutao

(School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract:** Mainstream method for scene reconstruction based on urban neural radiation field (NeRF) is to divide the entire scene into multiple regions. However, in unmanned aerial vehicle (UAV) aerial scenes, once the flight trajectory changes and the scene distribution is uneven, the segmentation efficiency of existing methods is low, making it difficult to handle the details of images taken at different heights of the same scene, resulting in poor rendering quality. Therefore, an end-to-end attention guidance framework is proposed. Firstly, utilizing camera pose for adaptive spatial partitioning to address the efficiency issue of non-uniform scene segmentation; Secondly, extracting and fusing the most relevant features of the target view from images at different heights to enhance details and solve the problem of high fidelity rendering in aerial scenes. The experiment shows that the proposed model achieved the best performance on both the urban scene datasets 56Leonardv and SCUTic, with a rendering speed improvement of over 65% compared to the most efficient methods currently available.

**Key Words:** adaptive spatial partitioning; feature fusion; attention-guided; NeRF model; urban scene

## 0 引言

大规模城市场景三维重建在数字化智慧城市建设方面具有重要的应用前景,由于传统三维重建方法需要在离

散3D体素或三角网格上操作,因此模型的分辨率和细节程度受到了一定的限制。目前,神经辐射场(Neural Radiation Field, NeRF)的方法在性能和质量上都超过了基于网格和体素的方法,在连续的3D空间中可生成具有任意分辨率和任意细节的模型,应用范围越来越广泛<sup>[1-7]</sup>。

收稿日期:2024-10-21

基金项目:四川省科技计划项目(2023YFG0304);四川省重大科技专项(2022ZDZX0008)

**作者简介:** 梁书凝(1998-),男,成都信息工程大学计算机学院硕士研究生,研究方向为三维重建;谭诗瀚(1986-),男,博士,成都信息工程大学计算机学院讲师,研究方向为计算机图形学、扩展现实;戈文一(1986-),男,博士,成都信息工程大学计算机学院讲师,研究方向为计算机图形学、扩展现实;钟娟(1989-),女,成都信息工程大学计算机学院硕士研究生,研究方向为三维重建;王录涛(1979-),男,博士,成都信息工程大学计算机学院副教授、硕士生导师,研究方向为机器视觉、嵌入式人工智能。本文通讯作者:王录涛。

在利用 NeRF 渲染大尺度场景时,因模型容量限制会导致重建细节表达不足、GPU 内存消耗过多,出现过度模糊的伪影和不完整的重建。为了解决该问题,一方面,Block-NeRF、Mega-NeRF、Switch-NeRF、BungeeNeRF 等主流方法在空间上将场景分解为多个单元,并在合并前为每个单元训练一个子 NeRF<sup>[8-11]</sup>;另一方面,Grid-NeRF、GP-NeRF 等将平面和网格特征与位置编码并行应用,以实现高效建模<sup>[12,13]</sup>。

然而,上述方法均基于分布均匀的大规模场景视图进行重建,在处理飞行轨迹多变的无人机航拍场景时效果欠佳,原因为 Mega-NeRF、Block-NeRF 主要使用 3D 采样距离或街道块来分解场景,不适合分布不均匀的无人机航拍场景。例如,Switch-NeRF 通过学习进行区域划分,需要一个专家网络确定每个采样点属于哪个区域,一次性输入所有图像信息占用了很高的 GPU 内存,当飞行轨迹多变时所需内存资源急剧增加。BungeeNeRF 利用高度分阶段训练,随着高度变化不可避免地需要处理大量的模型容量,占用了昂贵的训练时间,对有限的 GPU 计算资源而言不友

好。综上,这些方法在处理飞行轨迹多变的无人机捕获场景时存在两个问题:一方面,无人机飞行高度、方向变化时场景内的相机分布不均匀,上述方法在区域划分后视图可能差别过大,因此渲染效果不理想;另一方面,无人机在飞行变化过程中的视图存在显著的细节变化,现有模型难以捕获图像特征,因此渲染质量很差,同时变化也不可避免地使模型容量大量增加,延长了训练时间,需要大量的 GPU 计算资源。

为此,本文设计了一种端到端的注意力引导框架,如图 1(a)所示。首先,针对场景分布不均匀问题,利用无人机位姿聚类,将城市级场景划分为多个区域,为每个区域添加边界相机和相似相机,确保有足够多的相机能观察到该区域,以解决不均匀场景分割的效率问题,增强渲染效果;其次,针对细节变化引起的渲染质量问题,根据近距离摄像机捕获的区域构成远距离摄像机捕获的区域子空间;最后,基于注意力的特征融合方法,从不同高度的图像中提取潜在特征作为 NeRF 处理的场景先验信息,以增强模型对细节的处理能力,提升渲染精度。

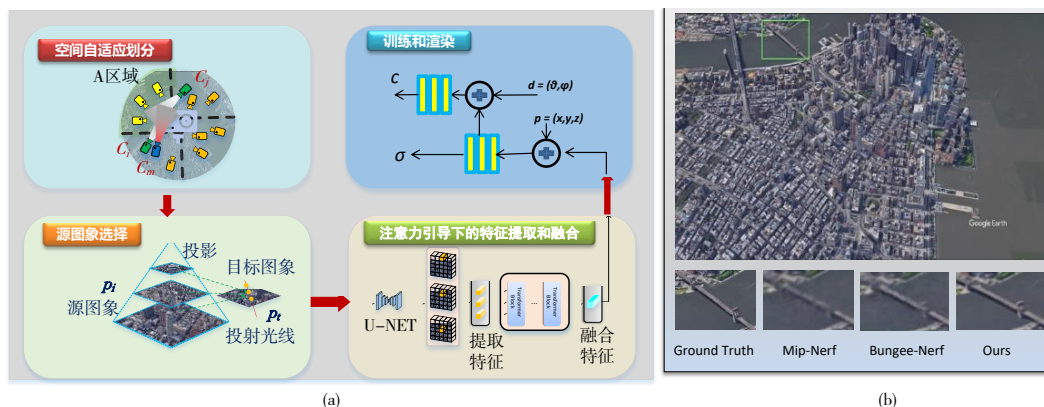


Fig. 1 Process flow and visualization results

图 1 处理流程与可视化结果

本文主要贡献为:

(1)设计了一种适用于大规模城市场景重建的端到端注意力引导框架。相较于当前主流大规模航拍渲染模型,在渲染速度上提升了近 3 倍,独特之处在于能灵活应对多尺度、不均匀的航拍数据集,且在适当的分割区域配置下仅需单个 GPU 即可完成任意规模航拍场景的渲染任务。

(2)提出一种自适应的空间划分与选择策略。基于相机位姿的相似性进行划分,在处理无人机捕捉的分布不均场景时既能显著提升渲染速度,还在渲染质量上优于现有模型。

(3)为了提升渲染质量,设计了高效的源图像筛选机制和基于注意力的特征融合技术。从不同高度的图像中有效提取关键特征,并将其作为 NeRF 场景的先验信息,增强了模型的细节处理能力。由图 1(b)渲染效果可见,本文方法相较于以往的大场景重建方法,在准确性、处理速度上均达到了业界领先水平。

## 1 相关工作

### 1.1 NeRF 及其扩展

NeRF 是一种革命性的 3D 重建技术,突破了传统 3D 重建方法基于离散 3D 体素或三角网格操作的限制。NeRF 利用多层感知器 (Multilayer Perceptron, MLPs) 在连续的 3D 空间中操作,将每个采样点的空间位置和视图方向沿着视图光线映射到颜色和密度上,生成具有任意分辨率和细节的 3D 模型,以减少计算复杂性,提升视点合成质量。

然而,NeRF 技术仍存在以下问题亟需解决:①NeRF 的训练和渲染过程相对较慢,限制了其在实际应用中的实时性能;②原始 NeRF 模型主要针对静态场景设计,对动态场景的处理能力有限;③NeRF 的泛化性能也面临挑战,需要大量的视角数据来训练一个高质量模型;④NeRF 在处理大型户外场景时也存在困难,因为这些场景通常包含更

复杂的几何结构和光照条件。

为了克服这些挑战并扩展 NeRF 的应用范围, 当前研究主要集中在两方面: 一方面, 拓展应用场景, 研究涵盖无界场景 (NeRF++、NeRF in the wild、Mip-NeRF 360)、动态场景 (Dynamic NeRF、ENeRF、DynIBaR、AligNeRF)、少镜头场景 (pixelNeRF、FreeNeRF、Sparse-to-Dense Depth Priors for NeRF、Depth-supervised NeRF、Sparse RGB-D NeRF for High-Quality Views) 及大型户外场景 (Block-NeRF、Mega-NeRF、Switch-NeRF、BungeeNeRF、基于神经辐射场的多尺度视图合成), 以满足日益增长的 3D 重建需求<sup>[14-26]</sup>; 另一方面, 提升重建效率, 采用显式表示来加速 NeRF 的收敛和推理过程<sup>[27]</sup>。

## 1.2 城市级 NeRF 重建与渲染

目前, NeRF 已从物体级扩展到城市级场景的三维建模。对于大规模场景, 一次处理所有图像需要很高的 GPU 内存, 数据量会超过 NeRF 的表达能力, 导致细节渲染模糊。为此, Block-NeRF 采用分治策略来处理大规模场景, 将一个城市的街景分解为多个区域, 每个区域都由一个单独的 NeRF 表示, 以提升纹理细节的渲染精度。例如, Mega-NeRF 根据 3D 距离将训练像素聚类成可由不同 NeRF 模型单独训练的部分。

然而, 当目标场景的规模和复杂性增加时, 这两种方法都不可避免地要均衡子 NeRF 的数量与每个子 NeRF 捕获区域所有精细细节的所需容量, 并且分解手工制作的场景也要依赖目标场景的先验知识。例如, Switch-NeRF 通过学习进行区域划分以获得良好的渲染效果, 但需要将所有数据一次性送到门控网络, 随着场景规模增加计算资源也会线性增加, 特别是当无人机场景分布不均匀时会对渲染结果产生不利影响。Bungee-NeRF 利用残差学习策略训练基于 NeRF 的模型渲染城市多尺度图集, 逐渐激活 NeRF 位置编码中的高频通道, 并随着训练深入展开更复杂的细节以在不同尺度上实现细节层次渲染, 但存在因尺度不一致导致的伪影、细节偏差等问题, 需要在输入图像间手动调整分割尺度, 增加额外的工作量。Grid-NeRF 在联合学习方案下集成了基于网格与基于 NeRF 的方法, 使用紧凑的多分辨率地面特征平面表示来粗略捕获场景, 以获取更清晰的几何形状和更精细的细节, 特别是在具有复杂几何和纹理的区域中能呈现更具细节的逼真新视角。GP-NeRF 构建了一个统一的学习框架, 兼容了广泛使用的分割模型和 NeRF, 促进了上下文感知的 3D 场景感知, 集成了正交二维高分辨率的三平面特征和三维哈希网格特征, 并利用 Transformer 进行场聚合和射线聚合以实现高效的场景表示。

虽然, 这些方法在无人机分布均匀的场景上性能较好, 但在处理分布不均匀的场景时表现不佳。本文针对无人机捕捉到的分布不均匀的大规模场景, 提出端到端注意力引导框架, 基于相机姿态进行自适应空间划分和选择,

直接从不同高度捕获的图像中提取更有利的场景先验, 以利用小的 MLP 进行场景重建。实验表明, 该方法在渲染速度上相较于传统模型提升较大, 无需手动调整输入视图来明确分割, 减少了数据集的准备时间。

## 2 研究方案

本文对于给定的分布不均匀的大规模场景数据, 通过 COLMAP 等软件获得相机的内在和外在矩阵, 如图 2 所示<sup>[28]</sup>。首先, 通过聚类分析空间分区来划分所有视图, 选择相邻区域的边界相机和相似相机以增强区域交界处的细节; 其次, 通过源图像选择方法在每个区域的目标视图中选择  $N$  个源图像, 并沿着投射光线的每个采样点投影回源图像, 通过基于注意力的特征提取和融合方法获得融合特征; 最后, 通过体渲染合成每条光线的颜色、密度, 以生成符合大小的图像。

### 2.1 空间分区与选择

由图 2 中空间自适应划分可见, 大型城市市场景信息主要在  $xy$  平面上。为便于识别边界相机, 对所有图像在  $xy$  平面上进行聚类, 将场景划分为  $N$  个区域  $area_i (i = 1, 2, \dots, N)$ , 并且  $N$  个区域在  $xy$  平面中存在中心位置点  $o_i (i = 1, 2, \dots, N)$ , 以自动分区无人机拍摄的分布不均匀场景。

通过这种方式划分后, 可能会有来自  $area_j$  区域的相机看到  $area_i$  区域, 而边界相机和边界相似相机的特点是靠近其他区域, 有可能观察到其他区域的场景, 因此可考察区域之间的边界相机与边界相似相机, 让他们与本区域相机一起训练, 使用足够多的能观察到第  $i$  个区域的相机来训练该区域的 NeRF, 以更好地重建区域交界处的细节。

#### 2.1.1 边界相机条件

边界相机的特点是在两个区域之间, 且该相机能观测到其他特定区域, 如图 3(a) 所示。通过计算相机之间的距离、相机方向与特定矢量之间的角度, 以确定相机观测范围和其在不同区域之间的边界性质。目前, 通过两个方面来判断一个相机是否为边界相机:

(1) 距离判断。设置一个阈值  $\gamma$  判断两个相机之间的距离是否足够近, 以至于一个相机可能是两个区域的边界相机。对于任意两个相机  $p_j$  (在  $area_j$  中) 和  $p_i$  (在  $area_i$  中), 如距离小于  $\alpha$  则  $p_j$  有可能是  $area_i$  与  $area_j$  之间的边界相机。

(2) 方向判断。相机的方向在其自身坐标系中表示为  $d_c = (0, 0, 1)$ , 通过相机的旋转矩阵  $R$  和平移向量  $t$ , 可将相机的方向转换到世界坐标系中, 如式 (1) 所示。

$$d_w = R * d_c + t \quad (1)$$

假设相机  $p_j$  指向区域  $area_i$  的中心  $o_i$  的矢量表示为  $a$ , 计算相机方向  $d_w$  与矢量  $a$  之间的角度  $\theta$ 。如果  $\cos\theta > 0$ , 说明相机方向与矢量  $a$  之间的夹角小于  $90^\circ$ , 即相机  $p_j$  可观测到  $area_i$  区域; 如果相机  $p_j$  与  $area_i$  中的相机  $p_i$  距离小于  $\alpha$ , 且  $d_w$  与指向  $area_i$  的矢量  $a$  之间的角度  $\theta$  满足  $\cos\theta > 0$ , 则确认

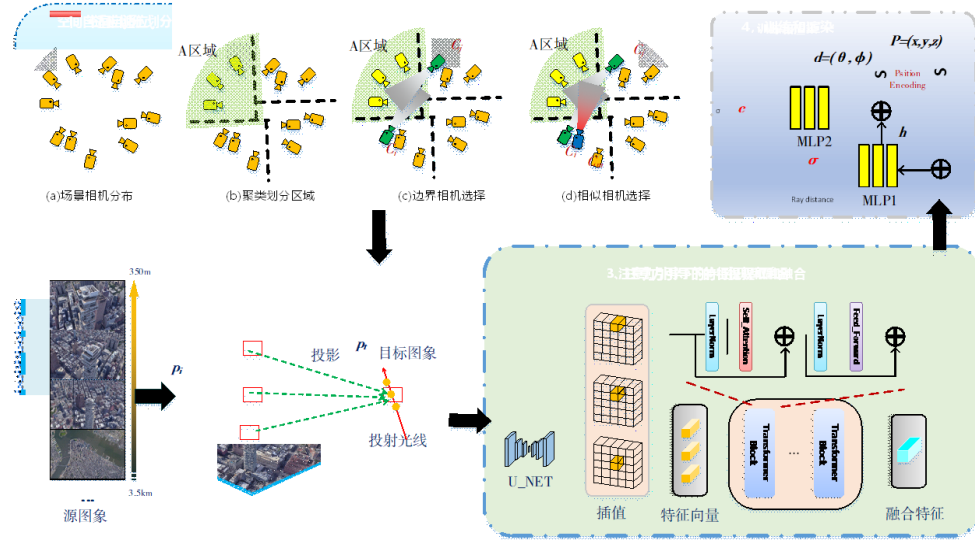


Fig. 2 Space adaptive partitioning and attention guided processing flow

图2 空间自适应划分与注意力引导处理流程

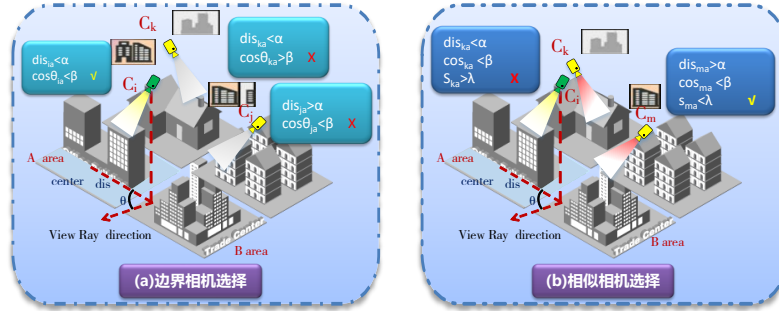


Fig. 3 Selection of boundary camera and similar camera

图3 边界相机和相似相机选择

相机  $p_j$  为  $area_i$  的边界相机。

### 2.1.2 相似相机条件

在  $area_j$  中存在一些相机,虽然不是边界相机但仍能查看  $area_i$ ,这些摄像机也需要进行区域重新定义,如图3(b)所示。对于摄像机  $p_i, p_j$ ,使用旋转  $R$ 、平移  $t$  来衡量他们的相似性。其中,  $R, t$  有助于保持摄像机空间位置和方向的相似性,他们之间的相似误差  $s$  为:

$$s = \|[R_i \ t_i] - [R_j \ t_j]\| \quad (2)$$

式中:  $s$  值越小说明两个摄像头越相似。

首先,计算每个边界相机与其他相机的相似误差,选择相似误差最小的  $n_s$  个相机作为这个边界相机的相似相机;其次,将相似相机与相应的边界相机分配到相同区域;最后,将边界相机和相似相机添加到一个区域,尽量全面地获取区域的图像信息,如算法1所示。

### 算法1 空间自适应分区和源图像选择算法

输入:图像位姿  $poses$ , K-Means 聚类数  $cluster\_num$ , 要选择的最近姿态的数量  $num\_select$ 。

输出:选择新聚类的源图像索引  $selected\_ids$ 。

1. function *Adaptive\_partitioning\_and\_image\_selection*
2. 相机 X 和 Y 坐标  $\leftarrow$  extract( $poses$ )

3. 分区类标签  $labels$ , 聚类区域中心  $cluster\_centers \leftarrow$  聚类  $K-Means(X, Y)$

4. FOR 相机数 DO
5. FOR 聚类区域数 DO
6. IF  $distance(\text{聚类区域相机}, \text{当前相机}) < \text{阈值 } \gamma$  THEN
7. 边界相机  $\leftarrow$  add list(当前相机)
8. END IF
9. END FOR
10. END FOR
11. FOR 边界相机数 DO
12.  $cac \leftarrow distance(\text{边界相机坐标}, \text{聚类中心 } cluster\_centers)$
13.  $cam\_xy\_dir \leftarrow$  投影 Project(相机方向, XY 平面)
14. 聚类的  $cosine\_cur \leftarrow cosine(cac \cdot cam\_xy\_dir, \|cac\|, \|cam\_xy\_dir\|)$
15. 类似计算与邻聚类的余弦值  $cosine\_neibor$
16. 边界相机区域  $cross\_label \leftarrow Addboundarycamera(cosine\_cur, cosine\_neibor)$
17. END FOR
18. FOR 聚类区域数 DO
19. delete( $cross\_label$  中的重复元素)
20. END FOR
21. FOR  $cross\_label$  DO

```

22. IF matrix THEN
23. list_distance ← angular_distance(matrix)
24. ELSE IF vector THEN
25. list_distance ← angular_distance(vector, 场景中心)
26. ELSE IF dist THEN
27. list_distance ← euclidean_metric(dist)
28. END IF
29. END FOR
30. sorted_ids ← sort(list_distance)
31. selected_ids ← min(sorted_ids, num_select)
32. RETURN selected_ids
33. END function

```

## 2.2 提取特征作为场景先验

目前, 城市场景重建是将整个场景编码为 MLP 层, 以增强场景表征, 但计算量和复杂度较高。由于自动空间划分后在同一区域的无人机航拍图像仍存在摄像机高度不同的现象, 高度变化中强烈的细节变化会导致渲染效果不佳。

为此, 本文基于近距离摄像机捕获的区域可构成远距离摄像机捕获的区域子空间这一原理, 采用一个图像选择系统来选择不同高度的少量图像以获得场景先验, 以粗略捕获场景几何形状、外观。通过这些场景先验作为引导信息在近距离视图中激活位置编码的高频通道, 或在远距离视图中激活低频通道, 基于自适应机制有效解决高度变化引起的细节变化造成的影响。

### 2.2.1 图像选择系统

由图 2 中源图像选择部分可知, 本文通过计算目标视图和各源视图之间的外部矩阵距离来判断两个视图的相似性, 从中选择值最小的  $N$  个图像作为源图像  $Select\_imgs$ , 原因为他们最接近目标视图, 可视范围高度重叠。

$$Select\_imgs = top\_N(\|P_t - P_i\|_2^2) \{P_i | i = 1, \dots, M\} \quad (3)$$

式中:  $P_i$  表示相机对训练数据集中第  $i$  张视图的外部矩阵;  $M$  为训练数据集中视图的个数;  $top\_N$  表示查找前  $N$  个最小值; 源图像选择如算法 1 结尾部分所示。

### 2.2.2 特征提取与融合

U-Net 是一种常用于图像分割和特征提取的卷积神经网络, 允许网络在多个尺度上捕获特征<sup>[29]</sup>。由图 2 中注意力引导下的特征提取和融合部分可知: 首先, 使用一个可训练的类 U-Net 网络处理源图像, 以提取一系列特征图  $\{F_i | i = 1, \dots, V\}$ ; 其次, 对于给定的 3D 点  $p \in R^3$ , 使用已知的相机内参矩阵将其投影到每个源图像的图像平面上, 得到图像平面坐标  $\pi_i(p)$ ; 再次, 在特征图上应用双线性插值提取特征向量, 即在每个源图像的特征图  $F_i$  上使用双线性插值在  $\pi_i(p)$  位置中提取特征向量, 允许不在整数像素位置的投影点也能精确定位、提取连续的特征值; 最后, 采样源图像  $S_i$  中  $\pi_i(p)$  处的 RGB 颜色值, 并连接提取的特征向量以形成最终特征向量  $\{f_i(p), i = 1, \dots, V\}$ 。

$$f_i(q) = \text{bilinear\_interp}(s_i, \pi_i(p)) + \text{bilinear\_interp}(F_i, \pi_i(p)) \quad (4)$$

式中:  $\oplus$  表示通道维度上的拼接;  $\pi_i(p)$  表示通过投影 3D 点  $p$  获得的图像平面坐标  $s_i$ 。

虽然这些高度重叠或全景源图像为目标视图提供了先验, 可粗略捕获场景几何形状和外观, 但引入了目标视图中不相关的信息。为了避免不相关信息的干扰, 本文使用一种基于注意力的特征融合方法, 核心思想是对于目标视图中的每个像素(或位置), 通过 Transformer 模型计算所有源图像特征向量与该像素的相关性, 相关性高的特征向量在融合过程中赋予更大权重, 从而确保融合后的特征向量与目标像素高度相关。Transformer 模型可输入多个特征向量, 并输出一个融合后的特征向量( $fatten(p)$ ), 不仅包含了源图像中的有用信息, 还通过注意力机制去除了不相关信息, 可精准表示目标视图中的每个像素。

$$fatten(p) = \text{Transformer}(f_1(p), f_2(p), \dots, f_v(p)) \quad (5)$$

在训练时, 将位置和方向编码、融合特征  $fatten(p_k)$  作为 MLP 的输入,  $fatten(p_k)$  可为 NeRF 的位置编码提供有关场景几何形状和采样位置外观的额外特征。在融合特征的指导下, NeRF 既能在大幅压缩的采样空间中更高效地工作, 还能在更少的采样点下捕捉更精细的细节, 提升渲染质量和效率。

## 2.3 训练与渲染

在每个空间区域选择图像的过程中, 采用位置编码将 3D 点坐标  $p_k = (x, y, z)$  与方向  $d_k = (\theta, \phi)$  映射到高维空间。本文将位置编码( $pos\_emb(p_k)$ ,  $pos\_emb(d_k)$ )和融合的特征  $fatten(p_k)$  作为 MLP1 的输入, 以获得第  $k$  个样本在射线上的颜色  $c_k$  和密度  $\sigma_k$ , 如式(6)所示。

$$\sigma_k, h_k = \text{MLP1}(\gamma(p_k), fatten(p_k)) \quad (6)$$

式中:  $h_k$  为一个中间特征, 起到连接位置信息与颜色预测的作用。

当 MLP2 预测颜色时, 不仅要考虑当前点的位置, 还要考虑通过 MLP1 学习到该点周围环境的复杂特征, 如式(7)所示。实验表明, MLP 与 NeRF 的结构类似, 但网络层更少, 能在更多有用信息的条件下缩短训练时间。

$$c_k = \text{MLP2}(\gamma(d_k), h_k) \quad (7)$$

## 3 实验结果与分析

### 3.1 数据集与指标

56Leonardv 数据集由现实世界不同高度的航空数据集组成, 跨越了约 0.3~3.5 km 的不同高度。SCUTic 数据集由无人机在空中不均匀飞行捕获的校园场景组成, 相机分布不均匀、相机距离大, 以测试不同方法的鲁棒性和泛化性。同时, 本文使用 PSNR、SSIM 和 LPIPS 等定量指标, 测试所提方法与先前工作方法的性能。其中, PSNR 测量两幅图像之间的均方误差; SSIM 侧重于结构相似性; LPIPS

测量感知相似性。

### 3.2 实验设置

在数据集与场景划分方面,使用56Leonardv、SCUTic两个数据集,将场景自动划分为4个区域,每个区域单独训练一个NeRF模型。在特征提取方面,每个区域先验源图像的数量为 $N=10$ ,使用类U-Net架构,以ResNet34为编码器,将两个上采样层、卷积和跳跃连接作为解码器,特征向量维度为35<sup>[30]</sup>。在网络架构方面,Feature融合模块采用两层Transformer block,MLP1共4层(每层64个特征维度),MLP2共两层(每层64个特征维度)。在射线采样与点设置方面,粗点和细点都设置为每条射线 $N_c=64$ , $N_f=64$ 。在优化与训练方面,使用Adam优化器,特征提取网络的初始学习率为 $10e^{-3}$ ,小型MLP的初始学习率为 $5 \times 10e^{-4}$ ,所有实验迭代20万次,每次迭代在单个RTX 3090 GPU上采样

2 048 条射线。

### 3.3 实验结果

#### 3.3.1 在56Leonardv数据集上的结果

图4(a)为56Leonardv空间分区划分结果,Bungee-NeRF将每个高度的相机分类为一类,使用从粗到精的方法从高到低训练NeRF,从高海拔地区到低海拔地区依次进行。由于图像纹理细节随海拔增加而减少,Bungee-NeRF会严重丢失纹理信息。实验表明,这种方法在最低高度、摄像机分布密集、相邻摄像机之间存在大量共享信息,将他们分组到一个类别可渲染出更准确的结果,但随着海拔增加场景范围扩大,建筑物的纹理细节会减少。为此,本文根据相机水平面位置分区降低图像复杂度,将不同高度的摄像机分到一个类别中,既提供了全局结构信息,又提供了纹理细节,还提升了渲染质量,如图5所示。

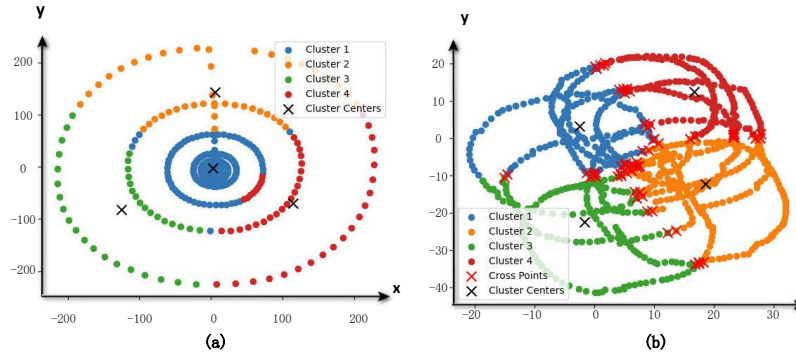


Fig. 4 Planar clustering partition results

图4 平面聚类分区结果

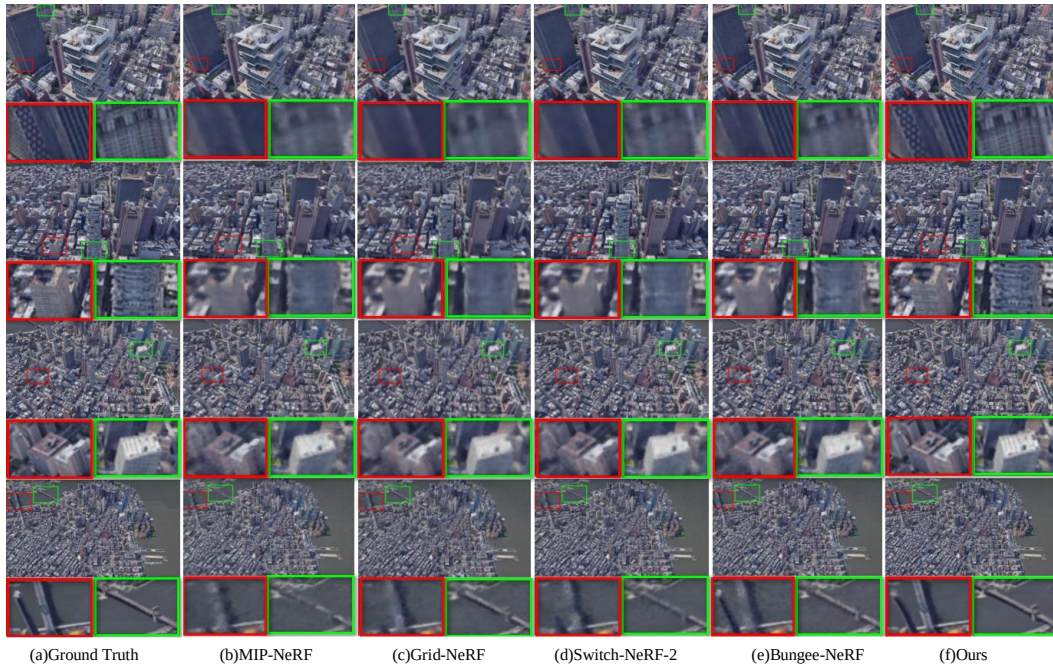


Fig. 5 Rendering results of the 56Leonardv dataset

图5 56Leonardv数据集的渲染结果

由表1可知,本文方法相较于Mip-NeRF,在PSNR、SSIM上分别提升1.965 dB和0.126 dB,在LPIPS上降低0.205 dB;相较于Grid-NeRF,在PSNR、SSIM上分别提升

1.537 dB和0.033 dB,在LPIPS上降低0.012 dB;相较于Switch-NeRF-2,在PSNR、SSIM上分别提升1.14 dB和0.273 dB,在LPIPS上降低0.347 dB。其中,Grid-NeRF需

要在 A100GPU 上进行训练, Switch-NeRF 需要多个 NVIDIA RTX 3090GPU 分布式并行训练数据, 对硬件环境要求较高。

此外, 本文方法只需训练 42 h, 相较于 Bungee-NeRF 节约了 65% 的时长, PSNR、SSIM 分别提升了 0.789 dB 和 0.02 dB, LPIPS 降低了 0.011 dB, 表明自适应空间划分与注意力引导对提升模型性能具有积极作用。

3.3.2 在 SCUTic 数据集上的结果

场景相机分布不均匀的数据集渲染结果如图 6 所示。由表 2 可知, Mip-NeRF 因不进行空间划分, PSNR 较低、渲染效果较差。Switch-NeRF 通过学习划分空间, 在投射光

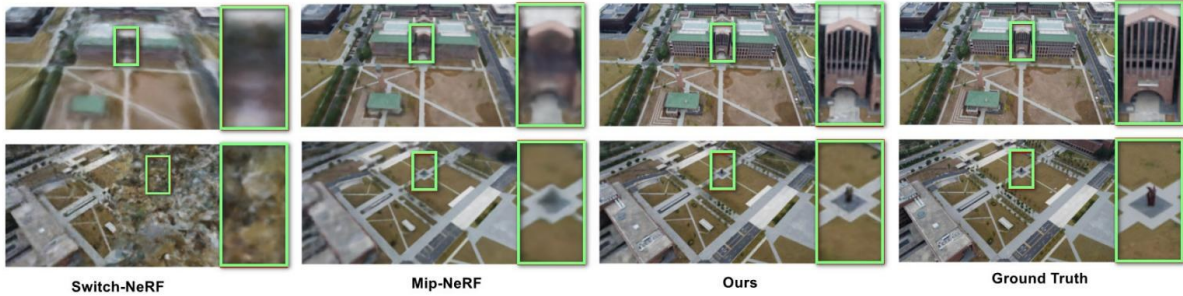


Fig. 6 Rendering results of SCUTic dataset

图 6 SCUTic 数据集渲染结果

Table 2 Performance comparison of different methods on SCUTic dataset

表 2 SCUTic 数据集上不同方法性能比较

| Method      | Partitioning | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|-------------|--------------|--------|--------|---------|
| Mip-NeRF    | 无            | 22.401 | 0.716  | 0.319   |
| Switch-NeRF | 学习           | 18.801 | 0.480  | 0.599   |
| Ours        | 自适应          | 27.602 | 0.846  | 0.109   |

本文方法根据无人机姿态划分空间, 以满足不同的飞行轨迹, 并为每个区域训练一个单独的 NeRF 来学习不同的空间信息。实验表明, 基于姿态相似性的空间选择算法可自适应将新视点映射到相应区域, 并使用该区域的 NeRF 进行渲染, 解决了信息差异问题。

图 5(b) 为 SCUTic 空间划分效果。由此可见, 在区域 2 中摄像机分布密集、信息量大, 因此分配给该区域的摄像机数量相对较少; 在区域 4 中摄像机稀疏、信息量小, 因此会分配更多的摄像机来增加该区域信息。通过合理分配摄像机和特征融合, 既提供了全局结构信息和纹理细节, 还提升了每个区域的渲染性能。针对分布不均匀的轨迹而言, 所提方法相较于 Switch-NeRF 在 PSNR、SSIM 上分别提升 8.801 dB、0.366 dB, 在 LPIPS 上降低 0.49 dB, 在应对不均匀的轨迹时性能更优。

3.4 消融研究

为了验证模型框架中各模块的性能, 设计了以下 4 组消融实验(见表 4): ①无空间划分。按聚类结果划分场景, 如果将场景均匀划分为多个区域时, PSNR、SSIM 分别降低 6.292 dB、0.242 dB, LPIPS 提升 0.339 dB; ②无边界相机选

Table 1 Performance of different methods on the 56Leonardv dataset

表 1 56Leonardv 数据集上不同方法的性能

| Method        | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---------------|--------|--------|---------|
| Mip-NeRF      | 23.337 | 0.709  | 0.354   |
| Grid-NeRF     | 23.765 | 0.802  | 0.161   |
| Switch-NeRF-2 | 24.162 | 0.562  | 0.496   |
| Bungee-NeRF   | 24.513 | 0.815  | 0.160   |
| Ours          | 25.302 | 0.835  | 0.149   |

线通过时需整合区域的不均匀信息, 原因为不均匀信息会导致不同区域之间的信息差异, 实验结果受空中飞行轨迹的影响较大。

择。设计边界相机条件、边界相机相似条件来选择特定区域的相机, 如果仅考虑区域相机, PSNR、SSIM 分别降低 3.231 dB、0.208 dB, LPIPS 提升 0.341 dB; ③无特征提取。为了验证源图像提供场景先验的有效性, 去除源图像选择方法( $N=0$ )后, 即注意力特征提取与融合模块没有输入, PSNR 降低 2.331 dB; ④无注意力引导(Maxpooling)。为了评估基于注意力的特征融合方法在捕获目标像素对应特征方面的有效性, 在插值特征向量中应用 Maxpooling, 实验表明在源图像数量相同的情况下, 本文模型相较于采用最大池化的模型 PSNR 降低 1.782 dB。

Table 3 Results of ablation experiment

表 3 消融实验结果

| Method             | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|--------------------|--------|--------|---------|
| 无空间分区              | 19.010 | 0.593  | 0.488   |
| 无边界选择              | 22.071 | 0.627  | 0.490   |
| 无特征提取( $N=0$ )     | 22.971 | 0.727  | 0.288   |
| 无注意力引导(Maxpooling) | 23.520 | 0.741  | 0.262   |
| 完整方法               | 25.302 | 0.835  | 0.149   |

4 结语

本文提出一种创新的端到端注意力引导框架, 处理飞行轨迹多变的无人机航拍城市场景。一方面, 通过自适应空间划分与选择策略有效解决了场景分布不均匀引起的应用效率低下问题; 另一方面, 基于注意力的特征融合, 从不同高度图像中提取、融合目标视图最相关的特征来增强细节, 提升渲染质量。

实验表明,所提模型相较于分解方法可更合理地分解复杂航拍场景,在场景合成上具有最先进的精度和速度水平,PSNR 值至少提升 3 dB,且渲染速度相较于目前最高效的方法仍提升 65% 以上。NeRF 方法不仅为大规模城市场景中的应用提供了新思路,还为实现高效、高保真的三维场景重建和渲染提供了有力支持。随着无人机技术不断发展和智慧城市建设的深入推进,将在未来展现出更广阔的应用前景。

#### 参考文献:

- [1] MILDENHALL B, SRINIVASAN P P, TANCİK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[J]. *Communications of the ACM*, 2021, 65(1): 99–106.
- [2] NIMIER-DAVID M, VICINI D, ZELTNER T, et al. Mitsuba 2: a retargetable forward and inverse renderer[J]. *ACM Transactions on Graphics*, 2019, 38(6): 1–17.
- [3] LIU S, LI T, CHEN W, et al. Soft rasterizer: a differentiable renderer for image-based 3D reasoning[C]// *Proceedings of the IEEE International Conference on Computer Vision*, 2019: 7708–7717.
- [4] ZHANG C, DONG Z, DOGGETT M, et al. Antithetic sampling for Monte Carlo differentiable rendering[J]. *ACM Transactions on Graphics*, 2021, 40(4): 1–12.
- [5] LOMBARDI S, SIMON T, SARAGIH J, et al. Neural volumes: learning dynamic renderable volumes from images[DB/OL]. <https://arxiv.org/abs/1906.07751>.
- [6] SITZMANN V, THIES J, HEIDE F, et al. Deepvoxels: learning persistent 3D feature embeddings[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019: 2437–2446.
- [7] MILDENHALL B, SRINIVASAN P P, ORTIZ-CAYON R, et al. Local light field fusion: practical view synthesis with prescriptive sampling guidelines[J]. *ACM Transactions on Graphics*, 2019, 38(4): 1–14.
- [8] TANCİK M, CASSER V, YAN X, et al. Block-NeRF: scalable large scene neural view synthesis[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 8248–8258.
- [9] TURKI H, RAMANAN D, SATYANARAYANAN M. Mega-NeRF: scalable construction of large-scale NeRFs for virtual fly-throughs[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 12922–12931.
- [10] MI Z X, XU D. Switch-NeRF: learning scene decomposition with mixture of experts for large-scale neural radiance fields[EB/OL]. <https://openreview.net/forum?id=PQ2zoIZqvm>.
- [11] XIANGLI Y, XU L, PAN X, et al. BungeeNeRF: progressive neural radiance field for extreme multi-scale scene rendering[DB/OL]. <https://arxiv.org/abs/2112.05504>.
- [12] XU L, XIANGLI Y, PENG S, et al. Grid-guided neural radiance fields for large urban scenes[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023: 8296–8306.
- [13] ZHANG Y, CHEN G, CUI S. Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features[DB/OL]. <https://arxiv.org/abs/2303.03003>.
- [14] ZHANG K, RIEGLER G, SNAVELY N, et al. NeRF++: analyzing and improving neural radiance fields[DB/OL]. <https://arxiv.org/abs/2010.07492>.
- [15] MARTIN-BRUALLA R, RADWAN N, SAJJADI M S M, et al. NeRF in the wild: neural radiance fields for unconstrained photo collections[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021: 7210–7219.
- [16] BARRON J T, MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: unbounded anti-aliased neural radiance fields[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 5470–5479.
- [17] PUMAROLA A, CORONA E, PONS-MOLL G, et al. D-NeRF: neural radiance fields for dynamic scenes[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021: 10318–10327.
- [18] LIN H, PENG S, XU Z, et al. Efficient neural radiance fields for interactive free-viewpoint video[DB/OL]. <https://arxiv.org/abs/2112.01517>.
- [19] LI Z, WANG Q, COLE F, et al. DynIBaR: neural dynamic image-based rendering[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023: 4273–4284.
- [20] JIANG Y, HEDMAN P, MILDENHALL B, et al. AligNeRF: high-fidelity neural radiance fields via alignment-aware training[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023: 46–55.
- [21] YU A, YE V, TANCİK M, et al. PixelNeRF: neural radiance fields from one or few images[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021: 4578–4587.
- [22] YANG J, PAVONE M, WANG Y. FreeNeRF: improving few-shot neural rendering with free frequency regularization[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023: 8254–8263.
- [23] ROESSLE B, BARRON J T, MILDENHALL B, et al. Dense depth priors for neural radiance fields from sparse input views[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 12892–12901.
- [24] DENG K, LIU A, ZHU J Y, et al. Depth-supervised NeRF: fewer views and faster training for free[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022: 12882–12891.
- [25] YUAN Y J, LAI Y K, HUANG Y H, et al. Neural radiance fields from sparse RGB-D images for high-quality view synthesis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 45(7): 8713–8728.
- [26] FAN T, YANG H, YIN W, et al. Research on multi-scale view synthesis based on neural radiance fields[J]. *Journal of Graphics*, 2023, 44(6): 1140–1148.  
范腾, 杨浩, 尹稳, 等. 基于神经辐射场的多尺度视图合成研究[J]. *图学报*, 2023, 44(6): 1140–1148.
- [27] HOU Y F, HUANG H S, FAN Q S, et al. A 3D reconstruction method based on improved multi-layer perceptron neural radiance fields[J]. *Laser & Optoelectronics Progress*, 2024, 61(4): 297–307.  
侯耀斐, 黄海松, 范青松, 等. 基于改进多层感知机的神经辐射场三维重建方法[J]. *激光与光电子学进展*, 2024, 61(4): 297–307.
- [28] PAN L, BARÁTH D, POLLEFEYS M, et al. Global structure—from-motion revisited[DB/OL]. <https://arxiv.org/abs/2407.20219>.
- [29] AZAD R, AGHDAM E K, RAULAND A, et al. Medical image segmentation review: the success of U-Net[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 10076–10095.
- [30] KUMAR S, YADAV A, VARSHNEY A. Comparative analysis of U-Net models using ResNet34, InceptionV3, and VGG16 for the processing of satellite images[C]// *Gwalior: IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*, 2024.

(责任编辑:刘嘉文)