中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

融合多源数据的大规模场景三维重建方法研究

作者姓名：高翔

指导教师：胡占义 研究员 中国科学院自动化研究所

申抒含 副研究员 中国科学院自动化研究所

学位类别：工学博士

学科专业：模式识别与智能系统

培养单位：中国科学院自动化研究所

2019 年 6 月

Large-scale Scene 3D Reconstruction Through Multi-source Data Fusion

A dissertation submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Pattern Recognition and Intelligent Systems

By
GAO Xiang
Supervisors:
Professor HU Zhanyi
Associate Professor SHEN Shuhan

Institute of Automation, Chinese Academy of Sciences

June, 2019

中国科学院大学
研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

受相机运动轨迹、拍摄姿态、环境光照与遮挡以及场景自身的几何结构和纹理分布情况的影响，如何提高大规模场景三维重建的完整性和效率长期以来一直是三维重建领域的一个挑战和目标。论文围绕大规模场景重建的完整性和高效性，重点研究了基于多源数据的重建方法，特别是融合航拍图像、地面图像以及激光扫描数据的方法。论文的主要工作包括以下四个方面：

1. 针对大规模建筑场景三维重建中基于地面图像建模完整度不够而基于航拍图像建模缺乏建筑立面细节的问题，提出了一种基于稠密点云的航拍与地面点云对齐的大规模建筑场景完整建模方法。该方法采用由粗到精的流程实现航拍与地面稠密点云的对齐。为提高点云对齐的精度与效率，该方法通过对地面稠密点云进行投影的方式实现航拍视角图像的合成。在点云对齐的过程中，该方法从图像选取、合成与匹配三方面进行了改进，使得合成的图像分布均匀，噪声较小，可得到更多的匹配内点。实验结果表明，该方法可有效地实现航拍与地面模型的精确、高效对齐。

2. 针对基于稠密点云投影的点云对齐方法效率较低，合成图像噪声大、有孔洞，且通过估计相似变换实现点云对齐无法处理基于图像的建模中的场景漂移等问题，提出了一种基于稀疏点云的航拍与地面点云融合方法。该方法采用基于稀疏网格诱导单应的方式合成航拍视角图像，并采用捆绑调整的方式实现航拍与地面点云融合，在一定程度上缓解了场景漂移问题。另外，该方法采用基于几何一致性检验和几何模型验证的方式对匹配外点进行过滤，实现了航拍图像与合成图像的有效匹配。实验结果表明，该方法在点云融合精度与效率方面优于其它对比方法。

3. 针对基于图像建模依赖环境因素，精度较低而基于激光数据建模灵活性低，成本高的问题，提出了一种融合图像与激光数据的精确、完整建模方法。该方法首先对场景进行图像采集并建模，基于图像建模结果，综合考虑场景结构复杂程度、纹理丰富程度以及扫描位置分布情况，自动规划激光扫描位置。之后，该方法通过激光点云投影合成图像，并与采集图像进行匹配。基于获取的图像与激光数据之间的跨数据类型特征匹配，采用由粗到细的流程，实现图像与激光数据的融合。实验结果表明，该方法能有效地实现图像与激光数据的精确融合。

4. 针对室内场景结构复杂、纹理不丰富，基于图像的建模结果不完整、不精

确的问题，提出了一种融合迷你飞行器与机器人数据的室内场景建模方法。该方法采用迷你飞行器采集图像构图，用于地面机器人路径规划并辅助机器人定位。为实现地面机器人的全局定位，该方法采用基于图割的方式合成机器人视角图像并将其与地面机器人采集的图像进行匹配。最后，通过融合迷你飞行器与地面机器人图像的方式，实现室内场景的精确、完整建模。实验结果表明，该方法可实现室内场景中地面机器人的精确定位以及场景的完整建模。

关键词：基于图像的三维建模，航拍与地面图像融合，图像与激光数据融合

Abstract

How to improve the scene completeness and computational efficiency has always been a key challenge and desired objective in large-scale 3D scene reconstruction community, due to various adverse factors, such as camera trajectory and view angle changes, environmental illumination condition and occlusions, as well as scene structure complexity and texture variation. This thesis focuses on how to use multi-source data, in particular, aerial images, ground images, and laser point cloud, to enhance the scene completeness and speed up the reconstruction process. The main results and contributions of the thesis are four-fold:

1. For large-scale architectural scene 3D reconstruction, the models generated from ground images are usually incomplete, while the models generated from aerial images lack fine details on the building facades. To tackle this problem, a dense point cloud based aerial and ground point cloud registration method for complete modeling is proposed, which goes in a coarse-to-fine way. In order to improve the accuracy and efficiency of point cloud registration, the proposed method synthesizes aerial-view image via ground dense point cloud projection. During point cloud registration, the proposed method makes several improvements in image selection, synthesis, and matching to generate evenly distributed synthetic images with low noise level and to obtain more point match inliers. Experimental results demonstrate that by the proposed method, accurate and efficient aerial and ground model registration could be achieved.

2. There are several drawbacks of the dense point cloud projection based point cloud registration method, for example, (1) relatively low efficiency, (2) synthetic image with high noise level and inevitable missing pixels, and (3) a similarity transformation for point cloud registration is incapable of modeling the scene drifting issue occurred in image based modeling. To deal with these issues, a sparse point cloud based aerial and ground point cloud merging method is proposed. In the proposed method, the aerial-view image is synthesized from the homographies induced by the sparse mesh, and the aerial and ground point clouds are merged via bundle

adjustment, which largely reduced the scene drifting problem. In addition, the proposed method filters the point match outliers between aerial and synthetic images via geometrical consistency check and geometrical model verification. Experimental results demonstrate that the proposed method performs better in point cloud merging accuracy and efficiency compared with other methods.

3. The models reconstructed from images are usually not accurate enough due to various factors, while the models generated from laser scans are of high cost and low flexibilities. In this work, an accurate and complete architectural scene modeling method by merging image and laser scans is proposed. The proposed method captures and models the scene using images at first. Based on the model generated from images, laser scanning locations are automatically planned by considering structural complexity and textural richness of the scene, and distribution of the scanning locations. Then, synthetic images are generated by projecting laser points, which are matched with the captured ones. Based on the cross-domain point matches, images and laser scans are merged by a coarse-to-fine scheme. Experimental results show that the proposed method could give accurate merging between images and laser scans.

4. Indoor scenes usually have complicated structures but texture paucity, it is hard to produce complete and accurate reconstructions by only image based modeling methods. This paper proposes a complete indoor scene modeling method using a mini drone and a robot. The proposed method uses aerial images captured by a mini drone to construct a global map, which is used to plan the moving path for robot and served as a global reference for robot localization. In order to localize the robot globally, the proposed method synthesizes ground-view image based on graph-cuts, which are then matched with the images captured by the robot on the ground. In the end, accurate and complete indoor scene models are achieved by merging aerial and ground images. Experimental results demonstrate that the proposed method is able to accurately localize the ground robot and completely model the indoor scene.

Keywords: Image Based 3D Modeling, Aerial and Ground Image Merging, Image and Laser Scan Merging

目 录

图形列表

表格列表

符号列表

缩写

ASIFT	affine-SIFT
AR	augmented reality
BA	bundle adjustment
BiCE	binary coherent edge descriptors
BIM	building information modeling
BoW	bag of words
CED	cumulative error distribution
FG	facet graph
FPS	frames per second
DLT	direct linear transform
FLAG	feature-based localization between air and ground
FLANN	fast library for approximate nearest neighbours
FPFH	fast point feature histogram
GCP	ground control point
GPS	global position system
ICP	iterative closest point
IMU	inertial measurement unit
IoU	intersection over union
IRLS	iteratively re-weighted least-squares
K-VLD	K-connected virtual line descriptor
LiDAR	light detection and ranging
LM	Levenberg-Marquardt
MAV	micro aerial vehicle
MVS	multiple-view stereo

NNDR	nearest neighbor distance ratio
NP	non-deterministic polynomial
PnP	perspective-n-point
PVR	probabilistic volumetric representation
QEM	quadric error metrics
RANSAC	random sample consensus
RMSE	root-mean-square error
RoI	region of interest
RoPS	rotational projection statistics
SfM	structure-from-motion
SFS	skeleton facet set
SI	spin images
SIFT	scale-invariant feature transform
SLAM	simultaneous localization and mapping
SURF	speeded-up robust features
UAV	unmanned aerial vehicle
UGV	unmanned ground vehicle
VR	virtual reality
WLCC	weighted largest connected component

第 1 章 绪论

1.1 研究背景及意义

基于图像的大规模场景三维重建一直是计算机视觉领域的一个重要的研究方向。近些年来，相关学者开展了一系列的研究工作并取得了许多代表性成果 [5–10]。随着重建算法与硬件设备性能的提升，算法的效率越来越高，重建场景的规模也越来越大。

在对重建结果进行评价时，重建精度与重建完整度是两个重要的评价指标 [11–13]。在现有的基于图像的大规模场景三维重建算法中，绝大部分的算法均聚焦于如何对相机位姿进行高精度估计并进而获得场景的高精度三维模型。而对于如何获取更加完整的三维模型，仅有为数不多的方法开展了相关研究 [14, 15]。上述两种方法分别基于曼哈顿世界假设 [14] 以及利用偏振图像 [15] 对弱纹理区域进行完整重建，尽管这两种方法在一些情况可取得较为完整的重建结果，然而在曼哈顿世界假设不成立的区域或者无法获取偏振图像的情况下，上述方法并不适用。在此，本文期望通过融合多源数据的方式进行大规模场景的三维重建，在重建过程中，兼顾重建的精度与完整度，以获取既精确又完整的重建结果。

影响大规模场景重建完整度的因素有很多，例如图像视角限制，场景结构复杂度、纹理丰富度等。基于图像的三维建模仅能重建图像中公共可见的场景，因此重建结果会较大程度上受图像拍摄视角的影响。例如，仅通过地面图像进行场景三维重建会丢失建筑屋顶信息，而仅通过航拍图像进行场景三维重建会缺少建筑立面细节。因此，本文的一个主要工作就是如何融合航拍与地面图像以实现大规模场景的精确、完整建模。上述问题的难点在于如何对尺度以及视角差异过大的航拍与地面图像进行特征匹配，本文采用的方式为利用地面数据（点云或图像）合成航拍视角图像，并与采集的航拍图像进行匹配。通过这种方式可以缓解航拍与地面图像间的尺度与视角差异，实现有效的图像匹配。

另外，由于基于图像的场景重建方法基于被动成像技术，待重建场景本身的一些特性，例如结构复杂度、纹理丰富度等也会影响场景重建的完整度。在结构复杂、纹理缺乏的区域，仅通过基于图像的三维建模方法会因为场景自遮挡或互遮挡以及缺乏特征匹配导致场景缺失，因此难以获得完整的场景重建结果。而相对而言，基于 LiDAR 数据的场景建模等主动成像技术对外界因素依赖较低，但这类

方法灵活性差，成本高。在此，本文的另一个工作就是以图像为主对场景进行完整覆盖，激光数据为辅对结构复杂、纹理缺乏区域进行补充，以获取精确、完整的场景重建结果。此方法通过投影激光点云合成图像与采集图像进行匹配，并通过全局优化实现图像与激光数据的融合。

除此之外，针对室内场景，融合迷你飞行器与地面机器人图像进行三维建模也可取得在精度与完整度方面优于仅采用单一来源图像的效果。由于在室内场景中地面机器人图像视场受限严重，且室内场景存在大量弱纹理区域，仅通过机器人图像对室内场景建模很容易因为误匹配或欠匹配导致场景飘移或场景缺失的情况。针对该问题，本文的最后一项工作就是采用迷你飞行器在室内采集图像并进行构图用于地面机器人的路径规划与全局定位，进而融合飞行器与机器人图像实现室内场景的精确、完整建模。

通过本文中的一系列融合多源数据的大规模场景三维重建方法，不仅能够获取更加完整的场景重建结果，在将数据融合并进行了全局优化以后，融合的数据反过来也可提升场景重建的精度。因此，本文中的场景重建方法兼顾重建的精度与完整度，可获取既精确又完整的重建结果。

1.2 研究现状

本节介绍了与本文相关工作的研究现状，共分为四类：（1）航拍与地面图像匹配；（2）航拍与地面模型对齐；（3）融合图像与激光数据建模（4）融合航拍与地面数据定位。各类工作的研究现状介绍如下。

1.2.1 航拍与地面图像匹配

为实现航拍与地面图像的匹配，一种直接的方式是通过采用常见的图像局部特征，如 SIFT[1]、SURF[16]、ASIFT[2] 等。然而，由于航拍与地面图像间的视角、尺度、光照通常情况下差异过大，采用上述特征难以达到预想的图像匹配效果。因此，为使局部描述子具备更强的鲁棒性，相关学者对描述子进行了一些规范化操作，提出了更为鲁棒的描述子，如 Root-SIFT[17]、BiCE[18] 等。上述方法虽然在鲁棒性方面有一定的提升，但是在匹配具有重复纹理特征的航拍与地面图像时，上述局部特征不能够得到正确的匹配。由于重复纹理导致的误匹配会对后续的模型对齐与模型融合造成影响。

另外，针对航拍与地面图像的匹配，有专门一类宽基线图像匹配方法。该类方

法基于视角校正，将待匹配图像均校正至正视视角 [19] 或者将地面图像校正至航拍图像视角 [4]，继而采用常规的，视角依赖的局部特征（如 SIFT）进行图像匹配。在进行图像校正时，可通过计算垂直与水平消隐点估计相机旋转矩阵的方式将图像校正至正视视角 [20]。在文献 [4] 中，Shan 等人采用将地面图像校正至航拍图像视角的方式，实现航拍与地面图像的匹配。首先，他们采用图像的 GPS 数据将重建的航拍与地面模型粗对齐至地理坐标系下；然后，根据通过 MVS 产生的稠密点云生成地面图像的深度图，采用基于深度的图像变换方式将地面图像校正至航拍图像视角；最后，利用 SIFT 特征进行合成图像与航拍图像的匹配，实现航拍与地面图像的匹配。

除此之外，还可通过匹配自相似图像块的方式进行航拍与地面图像之间的匹配 [21–23]。Wolff 等人 [23] 利用建筑物立面上的规律性结构，实现建筑物立面的航拍图像与街景图像的匹配。他们从图像中提取规则性图像块，并对候选匹配图像块的颜色、纹理及边缘上下文的相似性进行验证，获取正确匹配。然而，该方法只能确定从街景图像和航拍图像中分别提取的图像块是否属于同一个建筑物立面，而不能提供用于模型对齐的确切的两两匹配。

针对航拍与地面图像之间的大尺度差异问题，Zhou 等人 [24] 提出了一种基于尺度空间的尺度不变图像匹配方法。该方法基于尺度一致性原则 [25]，即特征匹配内点的尺度比值集中在图像尺度比值附近。他们通过基于 BoW 的编码方式估计图像尺度比。然后，根据估计的图像尺度比，将航拍与地面图像通过一种尺度已知的匹配方法进行匹配。该方法在图像间存在大尺度差异的情况下可提升特征匹配的精度、鲁棒性以及效率。

1.2.2 航拍与地面模型对齐

在文献 [4] 中，Shan 等人利用通过地面图像校正得到的航拍与地面图像的匹配点以及通过 GPS 初步对齐的航拍与地面模型，通过反投影得到模型上三维点的对应关系。之后，对得到的三维对应点采用 RANSAC[26] 估计上述模型间的相似变换矩阵，实现航拍与地面模型的对齐。

一些三维局部特征也可用于模型对齐。如 SI[27]、FPFH[3]、RoPS[28]。在一篇较为全面的算法评估文章中，Guo 等人 [29] 对一些三维局部特征在鲁棒性、拓展性和效率等方面进行了比较，他们的结论是 RoPS 性能最佳。然而，由于通过图像重建得到的三维模型通常噪声过大，使得上述三维局部特征并不适用于模型对

齐。另外，由于这两种模型并未较好地初始对齐，而且模型的精度、点密度及噪声程度通常也相差较大，ICP 算法 [30] 也并不适用。

另外，Wu 等人 [31] 提出了一种同时利用二维图像的纹理信息及三维模型的结构信息的新型局部特征。该特征基于图像校正，可用于模型对齐。该特征包括提取的特征点的纹理信息（SIFT 特征）及结构信息（空间位置、法向），具有视角不变性。因此，该特征在某些情况下可实现较为有效的模型对齐。然而，针对航拍与地面模型对齐问题，该特征效果较差。

Kaminsky 等人 [32] 提出了一种将地面模型对齐至垂直俯瞰航拍图像的方法。该方法基于图像边缘匹配并引入了自由空间约束，在求解最优对齐参数时，采用的是步进式搜索法。该方法在某些情况下能够实现地面模型与航拍图像的对齐。然而，当地面模型非立面信息或杂乱信息（如地面、植被等）过多时，该方法不能进行有效的边缘匹配且自由空间约束失效，继而得不到理想的对齐效果。

为给后续航拍与地面模型融合提供较好初值，Szomorú 等人采用 [33] 手动的方式实现了航拍与地面模型的对齐。他们通过地面控制点将航拍模型对齐至地理坐标系下，然后手动将地面模型对齐至航拍模型。另外，在文献 [33] 中，作者还提出了一种模型对齐精度的评价标准，他们通过考察两模型中相互最近邻点沿法向距离的分布情况来判断模型对齐的精度。

Zhou 等人 [34] 通过先对航拍与地面 MVS 模型模型重建得到表面网格，然后通过迭代的方式逐渐消除两网格之间的缝隙，以实现航拍与地面模型的精确对齐。该方法的基本思想与 ICP 算法类似，其关键在于如何建立两网格之间可靠的空间对应点。针对该问题，他们引入了一个名为 SFS 的集合用于表示网格上的局部平滑面片。网格间的对应点通过将航拍网格的 SFS 与整个地面网格投影至参考相机并进行深度比较获取。基于上述空间对应点，即可估计模型之间的相似变换 [35]。由于表面网格相对于 MVS 模型噪声程度更低，该方法可取得较方法 [4] 更为精确的模型对齐结果。

1.2.3 融合图像与激光数据建模

目前，存在一些方法通过融合图像与激光数据进行建模。然而，对于不同方法来说融合两种不同类型的数据的目的不尽相同。

一些工作通过利用底层特征（点或线）[36–38] 或高层特征（平面）[39] 将二维图像对齐至三维激光数据，用于为三维激光数据着色。基于对齐的二维图像与三

维激光数据，Li 等人 [40] 通过融合图像与激光数据，发挥两种数据各自的优势以获取完整、规则且带有纹理信息的城市场景中建筑物立面重建结果。

另外，在摄影测量学 [41]、计算机视觉 [11, 12] 与计算机图形学 [13] 领域，相关学者提出了一系列包含图像与激光数据的测评数据集用于测评重建算法。然而，在上述测评数据集中，激光数据通常用作真值，这种情况下图像与激光数据相对独立。另有一些方法 [42–44] 通过融合图像与激光数据以实现场景的完整重建。这些方法均基于三维模型对齐，通过借助 GCP[42] 或者 ICP 算法 [43, 44] 实现。

Frueh 等人 [45] 提出了一种将地面激光扫描仪获取的三维模型对齐至二维航拍图像与公路图的方法。据此以精确获取数据采集车位置的全局世界坐标并进而构建全局精确的三维模型。在他们的方法中，数据采集车搭载了两个二维地面激光扫描仪。一个竖直安装，其扫描平面垂直与行进方向，用于建筑物立面建模；另一个水平安装，其扫描平面平行于地面，通过站与站之间的匹配进行采集车相对定位。对于采集车绝对定位，他们提出了两种方法并进行了比较：（1）基于最大化互相关的定位，（2）基于马尔可夫-蒙特卡洛的定位 [46]。两种方法均将数字公路图和航拍图像与激光扫描数据进行结合实现定位。最后，该方法基于数据采集车的位置估计生成了城市场景的带有纹理的网格模型。

1.2.4 融合航拍与地面数据定位

对数据采集设备（相机或激光扫描仪）进行定位也可通过融合航拍与地面数据的方式进行。根据参考地图与采集设备数据类型的不同，融合航拍与地面数据的定位方法可分为三类：基于二维二维匹配的方法 [47–49]，基于二维三维匹配的方法 [50–52] 以及基于三维三维匹配的方法 [53–55]。

1.2.4.1 基于二维二维匹配的方法

Majdik 等人 [47] 尝试在存在 Google 街景图像数据集的地方对搭载相机的 MAV 进行全局定位。在进行航拍与地面图像匹配时，他们采取了类似 ASIFT[2] 的策略，即通过合成可能的视图以应对剧烈的视角变化。为加速匹配过程，他们采用了一种基于 FLANN[56] 与直方图投票方案 [57] 的候选图像匹配对的选取方式。最后，选取的匹配对通过一种基于图的匹配方法 K-VLD[58] 进行匹配。

Li 等人 [48] 提出了一种对地面图像进行地理定位的方法。该方法通过匹配地面图像中手绘的线段与正射图像中自动提取的线段实现。其中地面图像通过提供

人工标注的水平线与如下两个假设进行正射校正：（1）相机焦距 f 已知；（2）相机光轴平行于地平面。他们通过发现线段的不确定性经投影变换进行了非线性放大，设计了一种针对线段的不确定性模型。通过利用该不确定性模型，经正射校正的地面图像通过倒角匹配 [59] 的方式与候选正射图像进行匹配。

Viswanathan 等人 [49] 通过匹配 UGV 采集的图像与卫星或高空飞行器采集的图像，实现 UGV 的定位。他们首先将 UGV 拍摄的全景图像进行变换以构造地面地图。然后，他们对不同描述子在不同地图尺寸与地形类型进行地面地图与卫星地图匹配的表现进行了评测。基于上述结果，他们将算法各步骤采用粒子滤波 [46] 融入到贝叶斯定位框架中。最后，他们通过比较该方法估计的位置与街景图像的基于 GPS 估计的位置证实他们的定位算法优于 Google 街景图像定位算法。

1.2.4.2 基于二维三维匹配的方法

Matei 等人 [50] 提出了一种在城市场景中匹配地面查询图像与通过航拍平台获取的 LiDAR 数据的方法。该 LiDAR 数据已对齐至地理坐标系。他们首选对原始的激光点云进行分割，将其分为地面，建筑屋顶与其它杂乱区域（其余点）。分割的屋顶组合得到建筑并进一步获取多面水密模型。然后，他们从屋顶以及建筑物立面上获取建筑物外轮廓，并将其用作鲁棒特征与从地面图像上获取的类似特征进行匹配。接着，他们将从航拍激光数据上提取的特征渲染至预设的，可充分覆盖感兴趣区域的渲染位置。最后，他们对激光数据上的渲染建筑物轮廓与人工标注的地面图像上的建筑物轮廓进行匹配，实现地面图像的定位。

Ozcanli 等人 [51] 提出一种将地面图像地理定位至大规模三维表面模型的方法。该模型由航拍 LiDAR 数据生成的 PVR[60] 形式的模型。为对地面图像进行地理定位，他们生成了一个包含位置与朝向的完整相机位姿假定集。该集合可涵盖可能的查询图像位姿。然后，他们对相对于相机位姿假定集的三维表面模型的可见性、深度以及其它属性进行了预先计算并通过一个高效索引方法进行存储。最后，地面图像通过与上述步骤中生成的相机位姿假定集的属性图进行穷举式匹配实现地理定位。在他们的方法中，查询图像的语义分割结果通过人工标记给定。

Wang 等人 [52] 名为 FLAG 的方法。该方法通过地面图像与航拍图像上的特征计算机器人的全局位置更新。FLAG 借助垂直特征作为鲁棒的无描述特征，并通过匹配该特征与正射地图中已预知位置的垂直结构实现定位。他们利用双目立体相机检测垂直边并将其反投影至三维世界坐标系。然后，检测到的垂直边与全局

正射地图中预知的特征进行匹配。最后，地面机器人的位置通过粒子滤波框架 [46] 进行估计。

1.2.4.3 基于三维三维匹配的方法

Vandapel 等人 [53] 对利用空中高分辨率数据提升 UGV 在机器人定位与全局路径规划方面的表现进行了研究。该 UGV 在植被覆盖区域进行运动，他们采用的航拍与地面数据均为由 LiDAR 数据重建得到的三维表面网格。他们将由地面数据获取的地表模型对齐至航拍数据实现 UGV 的定位。在进行模型对齐之前，他们对航拍与地面数据中的植被进行了滤除以获取相对干净的地表模型。然后，（航拍与地面）地表模型通过 SI 特征 [27] 进行对齐。在模型对齐过程中，他们提出了通过引入位置约束与显著点选取策略以提升对齐效果。另外，在获取地面数据时，他们提出了一种全局路径规划的方法用于 UGV 自主导航。该路径规划方法利用高分辨率航拍数据并基于可通过性图实现。

Forster 等人 [54] 提出了一种在近距离下将一个 MAV 相对于一个地面机器人进行定位的方法。该问题通过对齐航拍与地面机器人传感器获取的三维地图进行求解。该方法中，航拍三维地图通过 MAV 上的单目相机采集的图像进行稠密重建获取；地面地图通过地面机器人上的深度传感器获取。他们通过如下两步实现航拍地面地图的对齐：（1）包含高程图与基于蒙特卡洛对齐的全局定位 [46] 以及（2）基于 ICP 的位姿优化 [61]。该方法定位精度高，实时性强。然而，该方法严重依赖于额外的传感器，如 IMU，以简化问题。

Surmann 等人 [55] 提出了一种用于多机器人协作的一致性三维地图构建方法。该方法将 UAV 与 UGV 获取的数据进行融合以构建一致性地图。他们通过 UAV 相机数据生成点云并将其与 UGV 上的二维激光扫描仪生成的点云进行融合。具体来说，该方法首先利用 UAV 采集的航拍视频通过 SfM 与 MVS 生成航拍稠密点云。然后，他们采用方法 [62] 对航拍与地面点云进行分割，得到平面结构。随后，UGV 的相对定位与绝对定位均通过基于平面块的对齐方法实现。需要注意的是，相对定位通过对齐 UGV 在不同位置生成的点云实现而绝对定位通过对齐航拍与地面点云实现。最后，他们通过全局优化融合航拍与地面点云以构建一致性地图。

1.3 基础知识

本文工作涉及到一些关于多视图几何与非线性估计的基础知识，为方便理解本文后续内容，在此对相关的基础知识进行简单介绍。

1.3.1 多视图几何

本节回顾了与基于图像的三维建模相关的多视图几何基本知识。本节首先介绍了单视图几何的一些基本概念，基于此，本节接着引入了与本文相关的一些两视图几何的内容。

1.3.1.1 单视图几何

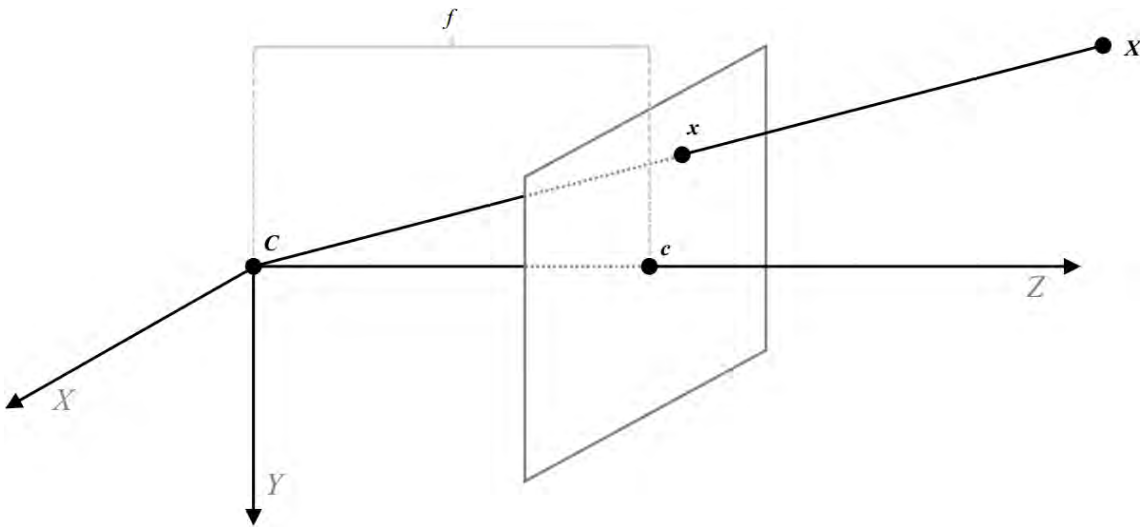

图 1.1: 针孔相机模型示意图。三维空间点 $\boldsymbol{X}$ 经针孔相机投影中心 $\boldsymbol{C}$，投影至像平面上的二维成像点，记为 $\boldsymbol{x}$。$\boldsymbol{c}$ 与 f 分别为主点与焦距。
Figure 1.1: Schematic diagram of the pinhole camera model. The scene point $\boldsymbol{X}$ is projected onto the image plane at $\boldsymbol{x}$ through the pinhole camera projection center $\boldsymbol{C}$. $\boldsymbol{c}$ and f are principal point and focal length of the camera, respectively.

针孔相机：本文中采用的相机模型为常用的透视投影模型，即由三维空间点 $\hat{\boldsymbol{X}} \in \mathbb{R}^3$ 射出的光线与二维像平面相交成像。当采用理想针孔相机模型时，所有的光线穿过同一个投影中心 $\boldsymbol{C} \in \mathbb{R}^3$。上述投影过程可通过公式如下表述：

$$\boldsymbol{x} \simeq \lambda[u \quad v \quad 1]^T = \boldsymbol{P}\boldsymbol{X} = \boldsymbol{K}[\boldsymbol{R} \quad \boldsymbol{T}]\boldsymbol{X} = \boldsymbol{K}[\boldsymbol{R} \quad -\boldsymbol{R}\boldsymbol{C}]\boldsymbol{X} \tag{1.1}$$

其中 $\boldsymbol{P}$ 为 3×4 矩阵，三维空间点 $\boldsymbol{X} \in \mathbb{P}^3$ 与二维成像点 $\boldsymbol{x} \in \mathbb{P}^2$ 均为齐次坐标形

式。3×3 旋转矩阵 $\boldsymbol{R} \in SO(3)$ 与平移向量 $\boldsymbol{T} \in \mathbb{R}^3$ 定义了世界坐标系到相机坐标系的欧氏变换。上述参数统称为相机外参数，而相机内参数由如下上三角矩阵表示：

$$\boldsymbol{K} = \begin{bmatrix} f & s & c_u \\ 0 & af & c_v \\ 0 & 0 & 1 \end{bmatrix} \tag{1.2}$$

其中 $\boldsymbol{c} = (c_u, c_v)^T \in \mathbb{R}^2$ 定义了主点位置，f 为以像素为单位的焦距长度，a 与 s 分别为非等量尺度因子与扭曲参数，分别用来表述像素长宽不同于不垂直的情况。针孔相机投影过程如图1.1所示。

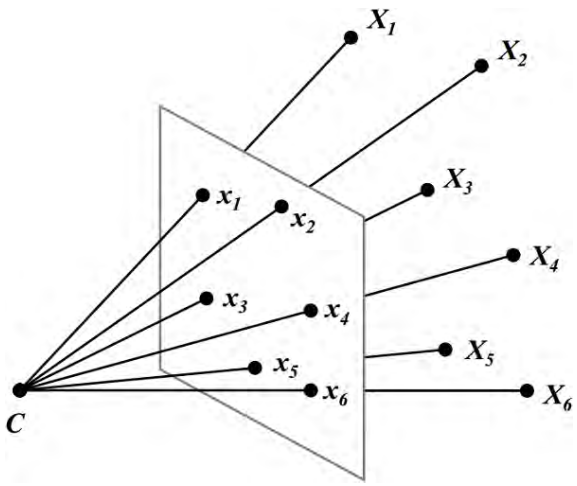

图 1.2: 通过 6 对二维三维对应点标定针孔相机示意图。
Figure 1.2: Schematic diagram of pinhole camera calibration from 6 2D-3D correspondences.

相机标定：相机标定的目的为恢复相机的内参数与/或外参数。给出二维图像观测点 $\boldsymbol{x}$ 与对应的三维空间点 $\boldsymbol{x}$，$\boldsymbol{P}$ 矩阵中的 12 个参数可通过式1.1定义的线性关系求解：

$$\boldsymbol{x} \simeq \boldsymbol{P}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{P}_1^T \\ \boldsymbol{P}_2^T \\ \boldsymbol{P}_3^T \end{bmatrix} \boldsymbol{X} \tag{1.3}$$

通过 DLT 将未知的尺度因子 λ 消去以获取如下的齐次线性方程组：

$$\begin{aligned} \boldsymbol{P}_3^T \boldsymbol{X} u &= \boldsymbol{P}_1^T \boldsymbol{X} \\ \boldsymbol{P}_3^T \boldsymbol{X} v &= \boldsymbol{P}_2^T \boldsymbol{X} \end{aligned} \quad \Rightarrow \quad \boldsymbol{0} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \\ \boldsymbol{P}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T & \boldsymbol{0}^T & -\boldsymbol{X}^T u \\ \boldsymbol{0}^T & \boldsymbol{X}^T & -\boldsymbol{X}^T v \end{bmatrix} \begin{bmatrix} \boldsymbol{P}_1 \\ \boldsymbol{P}_2 \\ \boldsymbol{P}_3 \end{bmatrix} \tag{1.4}$$

由于每对二维图像与三维空间对应点可提供 2 个约束，而上述方程组共含有 12 未知量，因此该方程组可通过至少 6 对对应点进行求解（见图1.2）。该方程组的直接线性变换解位于矩阵 $\boldsymbol{A}$ 的零空间中。假设 $\mathrm{SVD}(\boldsymbol{A}) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ 为矩阵 $\boldsymbol{A}$ 的奇异值分解，则式1.4的解为矩阵 $\boldsymbol{V}$ 的最后一列。另外，由于内参矩阵 $\boldsymbol{K}$ 为上三角矩阵，旋转矩阵 $\boldsymbol{R}$ 为正交矩阵，可以通过 RQ 分解由矩阵 $\boldsymbol{P}$ 获取相机的内外参数 [63]。

1.3.1.2 两视图几何

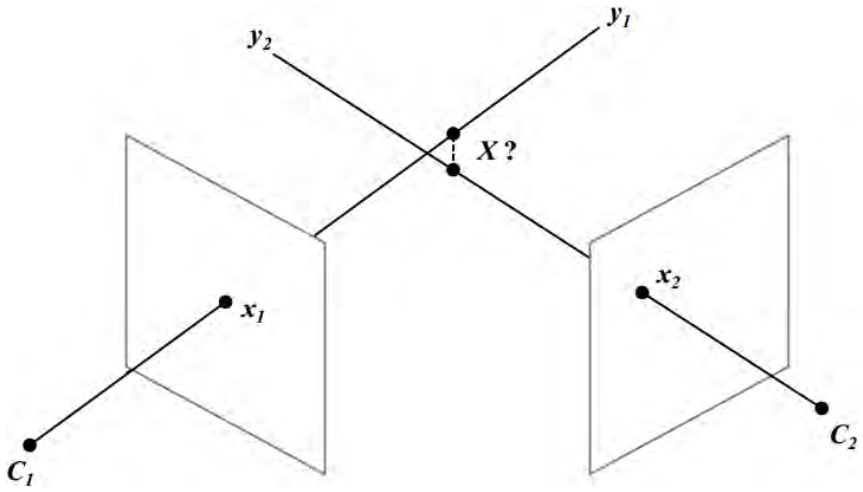

图 1.3: 空间点的两视图三角测量示意图。该过程通过对空间点 $\boldsymbol{X}$ 在图像中的观测点 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$ 对应的视线 $\boldsymbol{y}_1$ 与 $\boldsymbol{y}_2$ 进行相交实现。
Figure 1.3: Schematic diagram of two-view triangulation of a spatial point. The process is realized by intersection of the viewing rays, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, which correspond to the image observations, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, of the spatial point $\boldsymbol{X}$.

三角测量：在基于图像的三维建模中，最终目的是通过二维图像上的观测点 $\boldsymbol{x}$ 恢复三维场景结构 $\boldsymbol{X}$。对式1.1进行简单求逆是不可行的，这是由于任意两个沿视线 $\boldsymbol{y} = \boldsymbol{K}^{-1}\boldsymbol{x} \in \mathbb{P}^2$ 的点在射影空间中是相等的，即：

$$\boldsymbol{y}_1 = \lambda \boldsymbol{y}_2, \lambda \neq 0 \tag{1.5}$$

由于有三维向二维的投影过程中深度信息丢失了, 因此基于图像的三维建模中的关键挑战之一就是恢复未知的尺度因子 λ。换句话说，必须恢复沿像素点 $\boldsymbol{x}$ 对应的视线 $\boldsymbol{y}$，从相机光心位置 $\boldsymbol{C}$ 到空间点 $\boldsymbol{X}$ 的距离 λ。给出相机的内外参数以及一个图像点 $\boldsymbol{x}$ 对应的尺度因子 λ，该点对应的空间点可计算如下：

$$\bar{\boldsymbol{X}} = \lambda \boldsymbol{R}^T \boldsymbol{K}^{-1} \boldsymbol{x} + \boldsymbol{C} \tag{1.6}$$

而当深度 λ 未知时，空间点的位置可通过该空间点不同相机中的投影对应的视线相交确定（见图1.3）。上述视线相交过程称为三角测量。与式1.4中的相机标定类似，三角测量也可通过 DLT 对式1.1重新排列进行求解：

$$\begin{aligned} \boldsymbol{P}_3^T \boldsymbol{X} u &= \boldsymbol{P}_1^T \boldsymbol{X} \\ \boldsymbol{P}_3^T \boldsymbol{X} v &= \boldsymbol{P}_2^T \boldsymbol{X} \end{aligned} \quad \Rightarrow \quad \mathbf{0} = \begin{bmatrix} \boldsymbol{P}_3^T u - \boldsymbol{P}_1^T \\ \boldsymbol{P}_3^T v - \boldsymbol{P}_2^T \end{bmatrix} \boldsymbol{X} \tag{1.7}$$

上式即使在两视图情形也是超定的，因为此时共有 4 个线性等式而未知量共有 3 个。该额外的自由度源于三维空间中两视线通常不会恰好相交。这是由于图像观测点 $\boldsymbol{x}$ 上存在测量噪声且投影矩阵 $\boldsymbol{P}$ 估计的不够准确。另需注意的是，当两相机光心重合时上述方程组为奇异方程组。从几何上来讲，此时所有视线重合而在此视线上的任意点均为有效解。

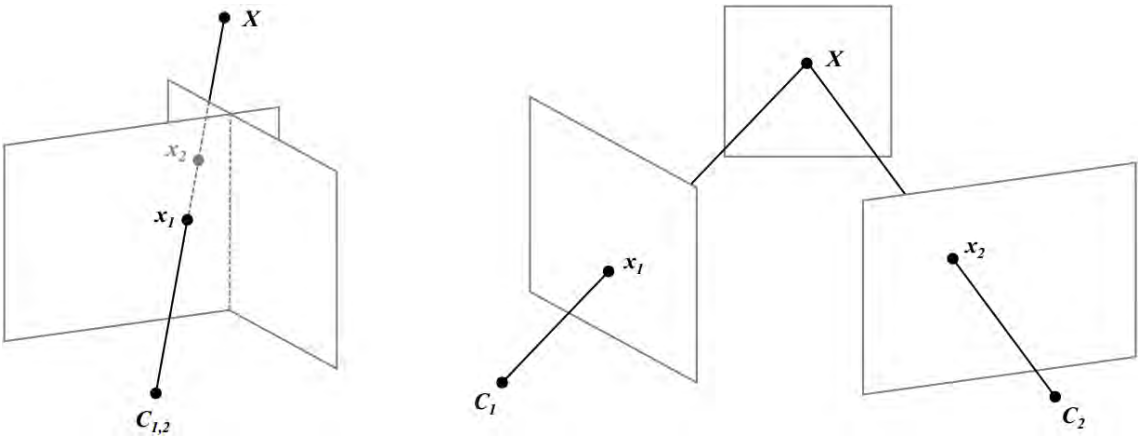

图 1.4: 单应矩阵描述平面上点的对应关系。它能将一个平面上的点（$\boldsymbol{x}_1$）映射到该点在另一个平面上的对应点（$\boldsymbol{x}_2$），因此可以描述相机纯旋转（左图）以及平面场景（右图）的两视图几何关系。

Figure 1.4: The homography describes the correspondence relationship between planes. It maps a point on one plane ($\boldsymbol{x}_1$) to its correspondence on another plane ($\boldsymbol{x}_2$) and thereby describes the two-view geometry for a purely rotating camera (left) and a planar scene (right).

单应变换：二维单应 h 为 $\mathbb{P}$ 内的射影变换，该变换可在射影空间中保持直线，即任意三点 $\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3$ 共线，当且仅当 $h(\boldsymbol{y}_1), h(\boldsymbol{y}_2), h(\boldsymbol{y}_3)$ 共线。换句话说，该单应将一平面上的点 $\boldsymbol{x}_1$ 映射到另一平面上的点 $\boldsymbol{x}_2$：

$$\boldsymbol{x}_2 \simeq \lambda_2 [u_2 \quad v_2 \quad 1]^T = h(\boldsymbol{x}_1) = \boldsymbol{H}\boldsymbol{x}_1 = \begin{bmatrix} \boldsymbol{H}_1^T \\ \boldsymbol{H}_2^T \\ \boldsymbol{H}_3^T \end{bmatrix} \boldsymbol{x}_1 = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \boldsymbol{x}_1 \tag{1.8}$$

由于射影歧义性 λ，单应 $\boldsymbol{H}$ 的自由度为 8。因此，$\boldsymbol{H}$ 可根据两幅图像中的 4 对对应点通过求解如下齐次方程组求解：

$$\begin{aligned}\boldsymbol{H}_3^T\boldsymbol{x}_1u_1 &= \boldsymbol{H}_1^T\boldsymbol{x}_1\\ \boldsymbol{H}_3^T\boldsymbol{x}_1v_1 &= \boldsymbol{H}_2^T\boldsymbol{x}_1\end{aligned} \quad \Rightarrow \quad \mathbf{0} = \begin{bmatrix}\boldsymbol{x}_1^T & \mathbf{0}^T & -\boldsymbol{x}_1^Tu_2\\ \mathbf{0}^T & \boldsymbol{x}_1^T & -\boldsymbol{x}_1^Tv_2\end{bmatrix} \tag{1.9}$$

由于单应可解释为两像平面之间的映射，它可以用于两种特殊配置下的两视图几何关系。在相机纯旋转的情形，投影中心与视线保持不变。如果在该相机的像平面上定义一个局部坐标系，可获取一个将点从第一个像平面映射到第二个像平面的单应（见图1.4）。注意，第二个像平面可以是场景中的任一平面，反之亦然。因此单应也可描述一个只看到平面场景的任意移动的相机的两视图关系：先构建将第一个像平面映射到场景平面的单应，再串联一个将场景平面映射回第二个像平面的单应（见图1.4）。

为不失一般性，将第一幅图像相机外参设为 $\boldsymbol{R}_1 = \boldsymbol{I}$ 与 $\boldsymbol{T}_1 = \mathbf{0}$。基于该假设，上文讨论的两种几何配置下的单应均可分解为：

$$\boldsymbol{H} = \boldsymbol{K}_2(\boldsymbol{R}_2 - \frac{\boldsymbol{T}_2\boldsymbol{N}^T}{d})\boldsymbol{K}_1^{-1} \tag{1.10}$$

在此，$\boldsymbol{N}$ 为场景平面的单位法向量，d 为场景平面到第一个相机投影中心的垂直距离。由于式1.8中固有的尺度不确定性以及式1.10中的 $\frac{\boldsymbol{T}_2}{d}$ 这一项的存在，单应通常不能恢复场景与相机运动的尺度。需要注意的是该特性对于基于图像的三维重建问题是固有的，在没有先验知识的前提下，无法恢复场景的尺度。

当相机进行纯旋转运动或者场景平面距相机无穷远时，单应简化为 $\boldsymbol{H} = \boldsymbol{K}_2\boldsymbol{R}_2\boldsymbol{K}_1^{-1}$。而对于相机已标定的情况（$\boldsymbol{K}_1$ 与 $\boldsymbol{K}_2$ 已知），单应简化为 $\tilde{\boldsymbol{H}} = \boldsymbol{R}_2 - \frac{\boldsymbol{T}_2\boldsymbol{N}^T}{d}$。仿射变换为单应的特殊情形，其用于表示焦距长度趋向于无穷时的正交相机，此时 $h_{31} = h_{32} = 0$，且 $h_{33} = 1$。

外极几何：上一节介绍了用于描述纯旋转或平面场景的单应，本节将介绍用于描述外极几何，它可用于描述一般场景下的相机非固定运动的两视图几何关系，此时两相机的投影中心 $\boldsymbol{C}_1$ 与 $\boldsymbol{C}_1$ 不同（见图1.5）。

为实现上述目标，首先定义外极点：另一个相机的投影中心在当前图像上的投影，即：

$$\boldsymbol{e}_1 = \boldsymbol{P}_1\boldsymbol{C}_2, \boldsymbol{e}_2 = \boldsymbol{P}_2\boldsymbol{C}_1 \tag{1.11}$$

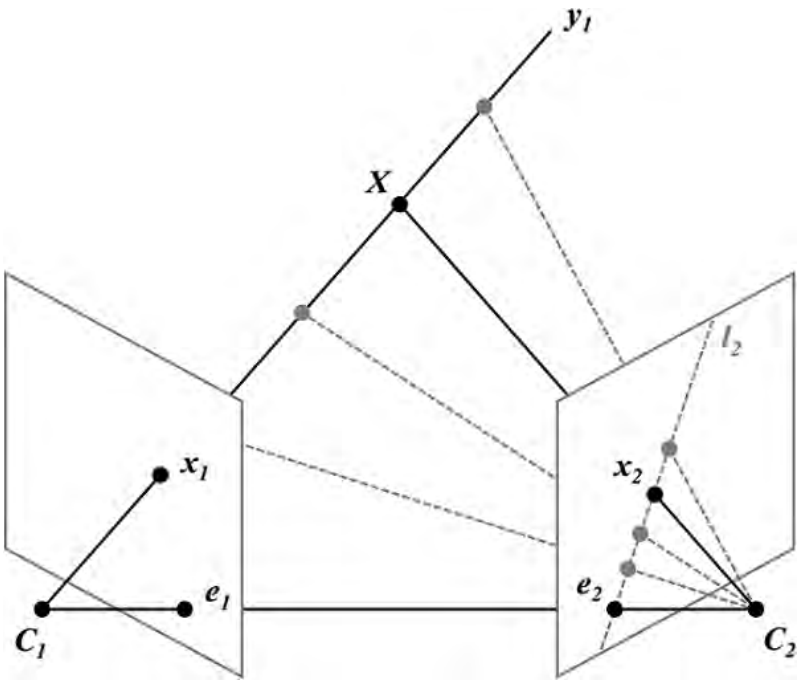

图 1.5: 基本矩阵描述任意场景中对应点满足的约束。它将一个图像上的点（$\boldsymbol{x}_1$）映射为另一个图像上的线（$\boldsymbol{l}_2$），且 $\boldsymbol{x}_1$ 的对应点 $\boldsymbol{x}_2$ 在 $\boldsymbol{l}_2$。因此，基本矩阵可用于描述一般场景下的相机一般运动的两视图几何关系。

Figure 1.5: The fundamental matrix describes the constraint the correspondence conform to in general scene. It maps a point ($\boldsymbol{x}_1$) from one image to a line ($\boldsymbol{l}_2$) in the other image, and the corresponding point $\boldsymbol{x}_2$ of $\boldsymbol{x}_1$ is on $\boldsymbol{l}_2$. As a result, the fundamental matrix describes the two-view geometry of a general scene under general camera motion.

对于任一三维空间点 $\boldsymbol{X}$，假设其在第一幅图像上的投影为 $\boldsymbol{x}_1$，可以通过外极点 $\boldsymbol{e}_1$ 与 $\boldsymbol{x}_1$ 构造外级线 $\boldsymbol{l}_1 = \boldsymbol{e}_1 \times \boldsymbol{x}_1$。将此外级线沿视线方向向场景进行延伸可得到外极平面 $\Pi = \boldsymbol{P}_1^+ \boldsymbol{l}_1$，其中 $\boldsymbol{P}_1^+$ 为 $\boldsymbol{P}_1$ 的伪逆。注意，所有的图像投影点对应的外极平面定义了绕两相机投影中心之间线段的平面簇。外极平面簇与像平面的交线定义了两个过外极点的外级线簇。外极平面 Π 与第二个相机的像平面交线即为第二幅图像中的外级线 $\boldsymbol{l}_2 = \boldsymbol{P}_2\Pi$。外极平面上的任意一点 $\boldsymbol{X}$ 在两幅图像上的投影分别位于两条外级线 $\boldsymbol{l}_1$ 与 $\boldsymbol{l}_1$ 上。上述外极约束强制 $\boldsymbol{X}$ 在第二幅图像上的投影 $\boldsymbol{x}_2$ 位于外级线 $\boldsymbol{l}_2$ 上，即：

$$0 = \boldsymbol{x}_2^T \boldsymbol{l}_2 = \boldsymbol{x}_2^T \boldsymbol{P}_2 \Pi = \boldsymbol{x}_2^T \boldsymbol{P}_2 \boldsymbol{P}_1^+ \boldsymbol{l}_1 = \boldsymbol{x}_2^T \boldsymbol{P}_2 \boldsymbol{P}_1^+ (\boldsymbol{e}_1 \times \boldsymbol{x}_1) = \boldsymbol{x}_2^T \boldsymbol{P}_2 \boldsymbol{P}_1^+ [\boldsymbol{e}_1]_\times \boldsymbol{x}_1 \quad (1.12)$$

其中，$[\boldsymbol{e}_1]_\times$ 为反对称矩阵，用于表示叉乘。

根据上述推导，外极几何关系可通过矩阵的形式定义如下：

$$\boldsymbol{F}=\boldsymbol{P}_2\boldsymbol{P}_1^{+}[\boldsymbol{e}_1]_{\times}=\boldsymbol{K}_2^{-T}[\boldsymbol{T}_{12}]_{\times}\boldsymbol{R}_{12}\boldsymbol{K}_1^{-1}=\begin{bmatrix}\boldsymbol{F}_1^T\\ \boldsymbol{F}_2^T\\ \boldsymbol{F}_3^T\end{bmatrix} \tag{1.13}$$

矩阵 $\boldsymbol{F}$ 称为基本矩阵，$\boldsymbol{R}_{12}=\boldsymbol{R}_2\boldsymbol{R}_1^T$ 与 $\boldsymbol{T}_{12}=\boldsymbol{T}_2-\boldsymbol{R}_{12}\boldsymbol{T}_1$ 分别为两相机之间的相对旋转与相对平移。基本矩阵将一幅图像中的点映射为另一幅图像的线。该矩阵秩为 2，自由度为 7，可以通过正常情况下的 7 对对应点（最小集）利用外极约束在相差一个尺度因子的意义下对其进行估计。上述最小解需要求解一个三次多项式，也可以通过对式1.12的齐次线性方程进行重排列，采用 8 对对应点对基本矩阵进行估计：

$$0=\boldsymbol{x}_1\otimes\boldsymbol{x}_2=[\boldsymbol{x}_1^Tu_2\quad\boldsymbol{x}_1^Tu_2\quad\boldsymbol{x}_1^T]\begin{bmatrix}\boldsymbol{F}_1\\ \boldsymbol{F}_2\\ \boldsymbol{F}_3\end{bmatrix} \tag{1.14}$$

上述方程可通过奇异值分解求解。上述方程组在一般配置下为超定方程组（8 个等式，7 个自由度），因此由于噪声的存在，在通常情况下秩为 2 的约束不能严格满足。这时，估计得到的基本矩阵将点映射获得的外级线不能恰好都交于同一个外极点。上述问题的最优解为找到一个与估计的基本矩阵最接近的秩为 2 的近似矩阵。另外，由于数值精度有限，在求解上述方程组时通常需要进行归一化操作 [64]。

本质矩阵为基本矩阵的相机已标定（内参矩阵 $\boldsymbol{K}_1$ 与 $\boldsymbol{K}_2$ 已知）的特殊情况。该矩阵仅有 5 个自由度，3 个为旋转，2 个为相差一个尺度因子意义下的平移。本质矩阵定义如下：

$$\boldsymbol{F}=\boldsymbol{K}_2^{-T}\boldsymbol{E}\boldsymbol{K}_1^{-1}\quad\Leftrightarrow\quad\boldsymbol{E}=\boldsymbol{K}_2^T\boldsymbol{F}\boldsymbol{K}_1=[\boldsymbol{T}_{12}]_{\times}\boldsymbol{R}_{12} \tag{1.15}$$

为估计本质矩阵，至少需要 5 对对应点（最小集）。与基本矩阵类似，本质矩阵也可通过 8 对对应点线性求解，而如果想采用最小配置的五点算法，需要求解一个十次多项式 [65]。本质矩阵可采用奇异值分解 $\mathrm{SVD}(\boldsymbol{E})=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ 得到 4 种可能的

相机相对位姿：

$$\begin{aligned} \boldsymbol{R}_{12}(\pm\pi) &= \boldsymbol{U}\boldsymbol{R}_z^T(\pm\frac{\pi}{2})\boldsymbol{V}^T \\ [\boldsymbol{T}_{12}]_\times(\pm\pi) &= \boldsymbol{U}\boldsymbol{R}_z(\pm\frac{\pi}{2})\boldsymbol{\Sigma}\boldsymbol{V}^T \end{aligned} \tag{1.16}$$

其中，$\boldsymbol{R}_z \in SO(3)$ 为绕 z 轴的旋转矩阵：

$$\boldsymbol{R}_z(\pm\frac{\pi}{2}) = \begin{bmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.17}$$

上述四个解仅有一个在几何上是有意义的，该解能使得通过三角测量得到的空间点位于两相机之前。与单应类似，本质矩阵也只能在相差一个尺度因子的意义下求取，因此通过分解本质矩阵得到的相机之间的相对平移也具有尺度不确定性。

1.3.2 非线性估计

上一节介绍了一些重建问题中的线性估计方法，这类方法计算效率高但是通常不足以获取精确且鲁棒的重建结果。上述情况的主要原因是带有噪声的观测值会产生不确定的估计结果。为缓解观测噪声对估计结果的影响，借助大量冗余的观测量以及对噪声模型正确建模都是十分关键的。绝大多数情况下，由于线性估计算法不能正确对噪声模型建模，无法获取最优解。接下来本节首先介绍了在高斯噪声模型假设下的最大似然的最优化算法，然后介绍了在观测量不满足高斯噪声模型时，如存在外点情况下 j 的鲁棒估计方法。

1.3.2.1 最优化算法

在进行非线性估计时，通常情况下都将求解问题的代价函数写成二次误差函数的形式，即获取待优化问题的最小二乘解。最小二乘意义下的代价函数通常形式如下：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{\theta})\|^2 \tag{1.18}$$

其中，其中 $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$ 为非线性代价函数，$\boldsymbol{\theta} \in \mathbb{R}^m$ 为待优化变量。通常情况下，形如上式的非线性代价函数有着许多局部极小值。由于对高度非线性的代价函数进行全局最小化求解十分困难，标准做法为对相对容易的原始代价函数的近

似进行迭代求解。举例来说，梯度下降法从初始估计 $\boldsymbol{\theta}_0^*$ 开始，通过沿着代价函数的局部线性近似的减小方向，即变量梯度方向:

$$\boldsymbol{g}(\boldsymbol{\theta}) = \nabla \frac{1}{2} \|\boldsymbol{f}(\boldsymbol{\theta})\|^2 = \boldsymbol{J}^T \boldsymbol{f}(\boldsymbol{\theta}) \tag{1.19}$$

对代价函数进行迭代最小化，其中，$\boldsymbol{J}_{ij} = \delta_j \boldsymbol{f}_i(\boldsymbol{\theta})$ 为 $\boldsymbol{f}$ 相对于 $\boldsymbol{\theta}$ 求偏导得到的 $n \times m$ 的雅可比矩阵。代价函数的一阶线性化表达为：

$$\boldsymbol{f}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) \approx \boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{J}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} \tag{1.20}$$

可通过如下表达式对其进行最小化：

$$\Delta\boldsymbol{\theta}^* = \arg\min_{\Delta\boldsymbol{\theta}} \frac{1}{2} \|\boldsymbol{f}(\boldsymbol{\theta}) + \boldsymbol{J}(\boldsymbol{\theta})\Delta\boldsymbol{\theta}\|^2 \tag{1.21}$$

通过上式可获取优化参数的迭代更新规则：由初始估计 $\boldsymbol{\theta}_0^*$ 出发，在第 t 步时 $\boldsymbol{\theta}_{t+1}^* = \boldsymbol{\theta}_t^* + \Delta\boldsymbol{\theta}_t^*$。注意，像是朴素梯度下降法之类的简单算法假设更新量 $\Delta\boldsymbol{\theta}$ 的步长为恒定值，因此这类方法经常会存在不收敛或者收敛速度极慢的问题。更加先进的基于置信域或者线性搜索的优化方法自适应地确定沿 $\Delta\boldsymbol{\theta}$ 的步长，因此这类方法有着更好的收敛特性。

对于绝大多数的基于几何的重建问题，尤其是对于 BA 来说，基于置信域的 LM 法 [66, 67] 是效率最高的优化算法。本质上讲，LM 法为梯度下降法与高斯牛顿法的结合。通过将式1.21展开并重排列，可获取高斯牛顿法的正则方程：

$$\boldsymbol{J}(\boldsymbol{\theta})^T \boldsymbol{J}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} = -\boldsymbol{J}(\boldsymbol{\theta})^T \boldsymbol{f}(\boldsymbol{\theta}) \tag{1.22}$$

而 LM 法的正则方程即为上述方程的改进版：

$$(\boldsymbol{J}(\boldsymbol{\theta})^T \boldsymbol{J}(\boldsymbol{\theta}) + \lambda \boldsymbol{D}(\boldsymbol{\theta}))\Delta\boldsymbol{\theta} = -\boldsymbol{J}(\boldsymbol{\theta})^T \boldsymbol{f}(\boldsymbol{\theta}) \tag{1.23}$$

其中，$\lambda > 0$ 为一个在梯度下降法与高斯牛顿法之间的阻尼系数，$m \times m$ 的对角阵 $\boldsymbol{D}(\boldsymbol{\theta})$ 用于衡量更新量 $\Delta\boldsymbol{\theta}$ 的质量，即当线性近似足够好时，该算法应当沿着梯度方向加大步长。为避免收敛速度过慢，Marquardt[67] 提出将增广海森用作 $\boldsymbol{D}(\boldsymbol{\theta})$，即 $\boldsymbol{D}(\boldsymbol{\theta}) = \mathrm{diag}(\boldsymbol{J}(\boldsymbol{\theta})^T \boldsymbol{J}(\boldsymbol{\theta})$。对于规模较大的优化问题，如果采用例如乔里斯基分

解的方法直接对线性方程 $\boldsymbol{J}(\boldsymbol{\theta})\Delta\boldsymbol{\theta} = -\boldsymbol{f}(\boldsymbol{\theta})$ 求解的话，通常来说效率较低。这是由于大规模优化问题的待优化参数与代价项数量过多。要想高效求解该类问题，通常需要根据 $\boldsymbol{J}(\boldsymbol{\theta})^T\boldsymbol{J}(\boldsymbol{\theta})$ 的系数结构设计求解方法。例如，在 BA 问题中，各幅图像都只能看见全部三维空间点中的较小的子集，即每个代价项仅与待优化参数中的较小子集相关，这样的话对应的雅可比矩阵也会十分稀疏。

1.3.2.2 鲁棒估计

上一节介绍了在高斯噪声模型假设下的最大似然估计算法，然而，当观测量不满足上述假设（存在外点）时，仍采用上述算法会导致偏差较大甚至错误的估计结果。出现这种情况的原因是观测外点会导致较大的误差项 $\boldsymbol{f}(\boldsymbol{\theta})$，而由于式1.21中的代价函数为二次方形式，该误差项会对估计结果产生较大影响。针对此问题，一种常用的做法是通过采用鲁棒损失函数对观测外点在代价函数中进行降权，以减小其对估计结果的影响。另外，还可以采用基于 RANSAC 的方法，通过迭代获取最大内点集的方式估计最优模型。

鲁棒损失函数：为减小式1.18中单个观测值的影响，将代价函数改写如下：

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_i \rho_i(\|\boldsymbol{f}_i(\boldsymbol{\theta})\|^2) \tag{1.24}$$

通过引入核函数 $\rho_i : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ 以减小单个误差项的影响。该做法的基本思想是通过选择合适的核函数 ρ_i，对非高斯误差项进行变换，使其符合高斯噪声模型。当估计结果接近最优解 $\boldsymbol{\theta}^*$ 时，对于内点来说（高斯测量值）误差项 $\boldsymbol{f}(\boldsymbol{\theta})$ 会较小，而对于外点来说（非高斯测量值）误差项 $\boldsymbol{f}(\boldsymbol{\theta})$ 会较大。因此，合适的核函数应当通过对较大的外点残差降权并对内点维持线性特性，使得误差由非高斯的重尾分布变换成高斯分布。当采用恒等核 $\rho_i(x) = x$ 时，上述等式变为原始的最小二乘问题，此时假设仅存在高斯误差项。下文将对几种用于鲁棒估计的最常见的核函数进行介绍。

IRLS 是一种根据当前最优估计对各误差项的权重进行迭代更新的算法。在当前第 t 次迭代中，各误差项的权重基于上一次迭代计算如下：

$$\rho_i(x) = \omega_i x, \omega_i = |\boldsymbol{f}_i(\boldsymbol{\theta})_{t-1}^*|^p \tag{1.25}$$

当 $p = 1$ 时，上式相当于最小绝对偏差。Huber 损失在误差低于特定的残差阈值 τ

时即为相对于观测量的二次损失，其定义如下：

$$\rho_i(x) = \begin{cases} x & \text{if} \quad x \leq \tau^2 \\ \tau(\sqrt{x} - \frac{\tau}{2}) & \text{if} \quad x > \tau^2 \end{cases} \tag{1.26}$$

Huber 损失非光滑且从函数形状来看尾巴较长，因此它仍然易受外点影响。Huber 损失有一种平滑近似 Pseudo-Huber 损失，定义如下：$\rho_i(x) = \tau^2(\sqrt{1 + \frac{x}{\tau^2}} - 1)$。另外，还有一种 Cauchy 损失，它平滑且更加鲁棒，通过如下方式计算：$\rho_i(x) = \log(1 + \frac{x}{\tau})$。

通常情况下，将鲁棒核引入非线性最小二乘问题中不会影响优化算法的使用。然而，此类经鲁棒化处理的问题的收敛特性通常会变得次最优。这是因为如果初始估计 $\boldsymbol{\theta}_0^*$ 距正确解较远的话，内点可能会被错误地降权，进而导致迭代优化算法难以收敛至正确解。文献 [68] 中提出通过将核函数提升至高维空间中可以克服上述问题。该方法通常有着更好的收敛特性，尤其是对于 BA 问题。

RANSAC：RANSAC 采用一种完全不同于上一节的思路处理鲁棒估计问题，其目标为估计一个模型，符合该模型的符合噪声假设的内点观测值数量最大化。更正式地说，RANSAC 的目标为最大化如下目标函数：

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{S} = \arg\max_{\boldsymbol{\theta}} \sum \rho(e_i^2), \rho(e^2) = \begin{cases} 1 & \text{if} \quad e^2 \leq \tau^2 \\ 0 & \text{if} \quad e^2 > \tau^2 \end{cases} \tag{1.27}$$

其中，e_i 为单个观测值的残差，$\mathcal{S}$ 为对应预先给定的内点距离阈值 τ 的内点观测值数量。举例说明，内点距离阈值可能是在估计相机内外参时以像素为单位的重投影误差。在高斯噪声模型假设下，内点距离阈值的一个典型选择为 $\tau = 3\sigma$，即认为满足噪声模型的观测量才可能为内点。

RANSAC 采取先假设后验证的策略来最大化目标函数 $\mathcal{S}$。该算法反复地随机选取 M 个观测量（最小集）并计算假设模型，然后通过在所有观测量中对内点计数以验证该模型。该算法的目标是在其各次迭代中，至少有一次选取的最小集合里的观测量均为内点。通过该最小集合，可以获得一个模型较好估计。有着最大内点支撑集 $\mathcal{S}^*$ 的模型即为最终的模型估计结果。

给出含 I 个内点的 N 个观测量，从中随机选取一个无外点的最小集合的概率

P 为：

$$\frac{\begin{pmatrix} I \\ M \end{pmatrix}}{\begin{pmatrix} N \\ M \end{pmatrix}} \approx \epsilon^M, \epsilon = \frac{I}{N} \tag{1.28}$$

RANSAC 的迭代次数 K 取决于三个变量 η, ϵ, M：

$$K = \frac{\log(1-\eta)}{\log(1-\epsilon^M)} \tag{1.29}$$

其中，η 为在 K 次随机选取中至少有一次选取的最小集合里的观测量均为内点的置信度。由上式可知，$K \sim \frac{1}{P}$ 通常情况下，由于事先并不知道内点比例 ϵ，先将其值设为最差情形，然后在迭代过程中一旦获取更多内点对应的模型，即对 ϵ 进行更新。在含外点比例较多的观测量中，采用尽可能少的观测量估计假设模型十分关键，因为这样的话能够提高选取的观测量子集中无外点的机会。需要注意的是，即使最小集合中的所有观测量均为内点，也不能保证能获取一个好的模型估计。这是由于采样集的观测量中含有噪声，且用于模型估计的观测冗余度不够。针对上述问题，一种解决方案是通过误差传播 [69] 对不确定性进行建模。另一种解决方案是采用局部优化的方式 [70]，通过最小集估计得到的模型对应的初始内点集对模型进行优化。然后将该优化过后的模型在全体观测量上进行验证，此时由于冗余度增加，该模型通常更加精确，因此对应更多内点。另外，还存在其它一些 RANSAC 的变种算法 [71]，有着更高的效率，更强的鲁棒性以及依赖更少的先验假设。

1.4 论文主要贡献

本文的主要贡献主要包括如下四个方面：

1. 针对基于地面图像建模完整度不够，缺失屋顶信息而基于航拍图像建模缺乏建筑立面细节的问题，提出了一种基于稠密点云的航拍与地面点云对齐的完整建模方法。该方法采用由粗到精的流程实现航拍与地面稠密点云的对齐。为提高点云对齐的精度与效率，该方法通过对地面稠密点云进行投影的方式实现航拍视角图像的合成。在点云对齐的过程中，该方法从图像选取、合成与匹配三方面进行了改进，使得合成的图像分布均匀，噪声较小，可得到更多的匹配内点。实验结果表明，该方法可有效地实现航拍与地面模型的精确、高效对齐。

2. 针对基于稠密点云投影的点云对齐方法效率较低，合成图像噪声大、有孔洞，且通过估计相似变换实现点云对齐无法处理基于图像的建模中的场景漂移等问题，提出了一种基于稀疏点云的航拍与地面点云融合方法。该方法采用基于稀疏网格诱导单应的方式合成航拍视角图像，并采用捆绑调整的方式实现航拍与地面点云融合，在一定程度上缓解了场景漂移问题。另外，该方法采用基于几何一致性检验和几何模型验证的方式对匹配外点进行过滤，实现了航拍图像与合成图像的有效匹配。实验结果表明，该方法在点云融合精度与效率方面优于其它对比方法。

3. 针对基于图像建模依赖环境因素，精度较低而基于激光数据建模灵活性低，成本高的问题，提出了一种融合图像与激光数据的精确、完整建模方法。该方法首先对场景进行图像并建模，基于图像建模结果，综合考虑场景结构复杂程度、纹理丰富程度以及扫描位置分布情况，自动规划激光扫描位置。之后，该方法通过激光点云投影合成图像，并与拍摄图像进行匹配。基于获取的图像与激光数据之间的跨数据类型特征匹配，采用由粗到细的流程，实现图像与激光数据的融合。实验结果表明，该方法能有效地实现图像与激光数据的精确融合。

4. 针对室内场景结构复杂、纹理缺乏，基于图像的建模结果不完整、不精确的问题，提出了一种融合迷你飞行器与机器人数据的室内建模方法。该方法采用迷你飞行器采集图像构图，用于地面机器人路径规划并辅助机器人定位。为实现地面机器人的全局定位，该方法采用基于图割的方式合成机器人视角图像并将其与地面机器人采集图像进行匹配。最后，该方法通过融合迷你飞行器与地面机器人图像的方式，实现室内场景的精确、完整建模。实验结果表明，该方法可实现室内场景中地面机器人的精确定位以及场景的完整建模。

1.5 论文结构安排

本文的结构安排如下：

第一章为绪论部分，介绍了融合多源数据进行大规模场景三维重建的研究背景及意义、研究现状，并介绍了本文中所用到的基础知识以及本文主要贡献。

第二章至第五章详细介绍了本文的四个主要工作，包括：（1）基于稠密点云的航拍与地面点云对齐方法，（2）基于稀疏点云的航拍与地面点云融合方法，（3）融合图像与激光数据的精确完整建模方法，（4）融合迷你飞行器与机器人数据的室内建模方法。

第六章对全文做出总结，并对未来工作进行展望。

第 2 章　基于稠密点云的航拍与地面模型对齐

2.1　引言

基于图像的大场景三维重建技术具有操作便捷、成本低廉、测量非接触等特点，在智能机器人、无人车、数字城市、文化遗产数字化、VR、AR 等领域均有广泛的应用。近年来，该项技术在重建精度，重建效率方面发展迅速 [10, 72–75]。

图 2.1: 由不同来源的图像重建得到的模型（稠密点云）。(a) 地面模型。(b) 航拍模型。
Figure 2.1: The reconstructed models (dense point clouds) from different image collections. (a) The ground model. (b) The aerial model.

为了保持待重建的大规模建筑场景的完整性，人们通常采用两种方式分别对场景进行图像采集及模型重建：(1) 通过手持数码相机拍摄图像重建地面模型；(2) 通过 UAV 搭载相机拍摄图像重建航拍模型。显然，通过这两种方式分别单独重建的模型（稠密点云）均存在各自的问题。如图2.1所示，地面模型虽具有丰富的细节特征，但由于地面图像视场、视角的限制，模型地面上存在孔洞且缺少屋顶信息（图2.1a）。反之，航拍模型虽更加完整，但由于航拍图像过低的空间分辨率，使得模型缺乏建筑物立面细节（图2.1b）。因此，可以通过将上述两种模型对齐的方式获取待重建建筑场景的较为精细、完整的三维模型。

本章提出了一种基于稠密点云投影的航拍与地面模型对齐方法。该方法将经由航拍与地面图像分别重建得到的航拍与地面三维模型通过由粗到精的过程对齐至同一坐标系下。首先，根据 GPS 信息将航拍与地面模型分别变换至地理坐标系下，实现模型的粗略对齐。其次，根据航拍与地面模型上的三维对应点估计模型间

的相似变换关系，实现模型的精细对齐。其中，三维对应点通过对航拍图像与将地面稠密点云投影至该航拍图像视角合成的图像进行匹配获取。本章方法的主要贡献总结如下：

- 本章提出了一种精确、高效的航拍与地面模型的对齐方法。
- 本章采用了整个地面模型而非单个地面图像进行航拍视角图像的合成。
- 本章方法在进行模型对齐时采用了三个关键步骤，包括图像选择、图像合成与图像匹配。
- 本章设计了一种用于参数评估以及方法对比的定量评价指标。

2.2 方法概述

为方便表述，本章所用符号如表所示。

表 2.1: 本章所用符号小结。根据不同情况，表格中符号的上标 M 可能为：IN 表示输入，CA 表示粗略对齐，FA 表示精细对齐，IS 表示初步航拍图像选取，FS 表示最终航拍图像选取，CS 表示当前航拍图像选取；下标 n 可能为 a 表示航拍相机/模型，g 表示地面相机/模型。表格中符号的 i 表示第 i 个相机。

Table 2.1: Summary of symbols and notations used in this chapter. Depending on the situation, the superscript M of the symbols in this table could be: IN for input, CA for coarse alignment, FA for fine alignment, IS for initial aerial view selection, FS for final aerial view selection and CS for current view selection; the subscript n could be: a for aerial camera/model and g for ground camera/model. i in this table denotes the i-th camera.

数据类型	符号	描述
相机	N_n^M	相机数量
	$I_{n(i)}^M$	图像
	$\boldsymbol{P}_{n(i)}^M$	投影矩阵
	$\{\boldsymbol{K}_{n(i)}^M, \boldsymbol{R}_{n(i)}^M, \boldsymbol{c}_{n(i)}^M\}$	内参、旋转矩阵以及光心位置
	$s_{n(i)}^M$	地面模型投影面积
	$r_{n(i)}^M$	地面模型投影面积比
	$\theta_{n(i)}^M$	航拍相机俯仰角
模型	$\mathcal{M}_n^M$	重建模型，即稠密点云
变换	$\{\mathbb{S}_n^M, \mathbb{R}_n^M, \mathbb{T}_n^M\}$	缩放、旋转以及平移

给出 N_a^{IN} 幅航拍图像 $\{I_{a(i)}^{IN} | i = 1, 2, \ldots, N_a^{IN}\}$ 及 N_g^{IN} 幅地面图像 $\{I_{g(j)}^{IN} | j = 1, 2, \ldots, N_g^{IN}\}$，本章方法采用 SfM[73] 获取相机参数（内参 $\boldsymbol{K}$ 及外参 $\boldsymbol{R}, \boldsymbol{c}$），将其分

别记为 $\{\boldsymbol{K}_{a(i)}^{IN}, \boldsymbol{R}_{a(i)}^{IN}, \boldsymbol{c}_{a(i)}^{IN} | i = 1, 2, \ldots, N_a^{IN}\}$ 及 $\{\boldsymbol{K}_{g(j)}^{IN}, \boldsymbol{R}_{g(j)}^{IN}, \boldsymbol{c}_{g(j)}^{IN} | j = 1, 2, \ldots, N_g^{IN}\}$。然后，本章方法采用 MVS[75] 获取航拍模型 $\mathcal{M}_a^{IN}$ 及地面模型 $\mathcal{M}_g^{IN}$。这里的模型指的是通过 MVS 获取的带有颜色及法向信息的稠密点云。另外，本章方法假设可以获得航拍与地面相机的 GPS 信息，即它们的地理坐标 $\{\boldsymbol{c}_{a(i)}^{GPS} | i = 1, 2, \ldots, N_a^{IN}\}$ 及 $\{\boldsymbol{c}_{g(j)}^{GPS} | j = 1, 2, \ldots, N_g^{IN}\}$。该信息可通过如下两种方式获取：

- 载入相机内置 GPS 信息；
- 由 GCP 转换得到，GCP 的地理坐标由差分 GPS 测得，用于高精度测量。

本章方法的输入为（1）航拍图像 $\{I_{a(i)}^{IN}\}$（无需地面图像 $\{I_{g(j)}^{IN}\}$）；（2）航拍与地面相机参数 $\{\boldsymbol{K}_{a(i)}^{IN}, \boldsymbol{R}_{a(i)}^{IN}, \boldsymbol{c}_{a(i)}^{IN}\}$，$\{\boldsymbol{K}_{g(j)}^{IN}, \boldsymbol{R}_{g(j)}^{IN}, \boldsymbol{c}_{g(j)}^{IN}\}$；（3）航拍与地面模型 $\mathcal{M}_a^{IN}$，$\mathcal{M}_g^{IN}$ 以及（4）航拍与地面相机的 GPS 信息 $\{\boldsymbol{c}_{a(i)}^{GPS}\}$，$\{\boldsymbol{c}_{g(j)}^{GPS}\}$。本章方法的最终输出为经过精确对齐的航拍与地面模型 $\mathcal{M}_a^{FA}$，$\mathcal{M}_g^{FA}$。

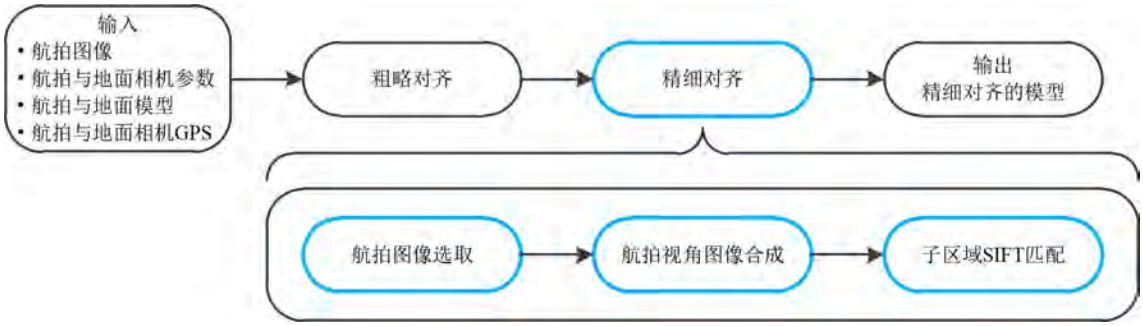

图 2.2: 本章方法流程图。
Figure 2.2: Pipeline of the proposed method in this chapter.

本章中的航拍与地面模型对齐方法分两步进行：首先根据航拍与地面相机的 GPS 信息将航拍与地面模型分别变换至地理坐标系下，实现航拍与地面模型粗略对齐；然后通过如下三个关键步骤实现地面与航拍模型精细对齐：航拍图像选取，航拍视角图像合成以及子区域 SIFT 图像匹配。本章方法的流程图如图2.2所示。

2.3 粗略对齐

在进行航拍与地面模型的粗略对齐时，由于航拍与地面模型均采用度量重建的方式获得，使得航拍与地面模型分别在各自的局部坐标系下。因此，为统一航拍

与地面模型的尺度，首先需要将模型变换至地理坐标系下：

$$\begin{aligned}\mathcal{M}_a^{CA} &= \mathbb{S}_a^{CA}\mathbb{R}_a^{CA}\mathcal{M}_a^{IN} + \mathbb{T}_a^{CA} \\ \mathcal{M}_g^{CA} &= \mathbb{S}_g^{CA}\mathbb{R}_g^{CA}\mathcal{M}_g^{IN} + \mathbb{T}_g^{CA}\end{aligned} \tag{2.1}$$

其中，$\mathcal{M}_a^{CA}$，$\mathcal{M}_g^{CA}$ 分别为经过粗略对齐后，处于地理坐标系下的航拍与地面模型；$\{\mathbb{S}_a^{CA}, \mathbb{R}_a^{CA}, \mathbb{T}_a^{CA}\}$，$\{\mathbb{S}_g^{CA}, \mathbb{R}_g^{CA}, \mathbb{T}_g^{CA}\}$ 分别为航拍与地面模型局部坐标系与地理坐标系之间的相似变换。$\mathbb{S}$，$\mathbb{R}$，$\mathbb{T}$ 分别为缩放，旋转及平移。上述相似变换可根据航拍与地面相机在局部坐标系与地理坐标系中的位置坐标，通过 RANSAC[26] 进行估计。以航拍模型为例，分别给出航拍相机在局部坐标系与地理坐标系下的坐标 $\{\boldsymbol{c}_{a(i)}^{IN}\}$ 及 $\{\boldsymbol{c}_{a(i)}^{GPS}\}$，本章方法从其中随机选择用于三维到三维点集对齐的最小航拍相机集合（大小为 3）。然后，采用最小二乘法 [35] 估计它们之间的相似变换，进而得到对应该相似变换的内点集。其中，用于求取内点的阈值在此设为 $1m$。上述过程重复 500 次以获取内点数量最多的最大一致集。最后，用于航拍模型粗略对齐的相似变换 $\{\mathbb{S}_a^{CA}, \mathbb{R}_a^{CA}, \mathbb{T}_a^{CA}\}$ 通过获取的最大一致集采用最小二乘法 [35] 估计得到。用于地面模型粗略对齐的相似变换 $\{\mathbb{S}_g^{CA}, \mathbb{R}_g^{CA}, \mathbb{T}_g^{CA}\}$ 也是采用相同的方式求取。

另外，由于下文中要用到航拍相机的位姿信息，而航拍相机的位姿信息需要与经粗略对齐的航拍模型保持一致。因此，航拍相机外参数也需要利用估计得到的相似变换粗略对齐至地理坐标系下同时保持其内参数不变：

$$\begin{aligned}&\boldsymbol{K}_{a(i)}^{CA} = \boldsymbol{K}_{a(i)}^{IN} \\ &\boldsymbol{R}_{a(i)}^{CA} = \boldsymbol{R}_{a(i)}^{IN}(\mathbb{R}_a^{CA})^T \qquad ,(i = 1, 2, \ldots, N_a^{IN}) \\ &\boldsymbol{c}_{a(i)}^{CA} = \mathbb{S}_a^{CA}\mathbb{R}_a^{CA}\boldsymbol{c}_{a(i)}^{IN} + \mathbb{T}_a^{CA}\end{aligned} \tag{2.2}$$

其中，$\{\boldsymbol{K}_{a(i)}^{CA}, \boldsymbol{R}_{a(i)}^{CA}, \boldsymbol{c}_{a(i)}^{CA}\}$ 为经过粗略对齐的航拍相机参数。

通常情况下，对于后续的精细对齐来说，仅根据 GPS 得到的粗略对齐结果已经足够精确。然而，在某些情况下，上述粗略对齐的结果不足以为后续的精细对齐提供较好的初值。例如 GPS 信息噪声过大或者航拍与地面相机的 GPS 信息来源不同。后一种情况主要指的是航拍模型根据由差分 GPS 精确测量的 GCP 的地理坐标进行粗略对齐而地面模型由相机内置 GPS 信息进行粗略对齐。值得注意的是，这种情况在航拍与地面联合建模中比较常见，这是因为相对于地面图像，在航拍图像中辨识 GCP 容易得多。

由于航拍相机的视场及分布区域远大于地面相机，因此航拍模型覆盖的场景区域也远大于地面模型。在此，本章方法首先在 $\mathcal{M}_a^{CA}$ 中人工选取与 $\mathcal{M}_g^{CA}$ 重合的子模型，并将其记为 $\mathcal{M}_{a\cap g}^{CA}$。然后，本章方法分别获取 $\mathcal{M}_{a\cap g}^{CA}$ 与 $\mathcal{M}_g^{CA}$ 的中心。接下来，$\mathcal{M}_{a\cap g}^{CA}$ 与 $\mathcal{M}_g^{CA}$ 相对于地理坐标系的局部旋转通过对它们的协方差矩阵进行特征值分解得到。另外，由于航拍与地面模型均已经粗略对齐至地理坐标系，它们有着相近的尺度。因此，粗略对齐的结果可以根据已获取的模型中心及局部旋转，通过将 $\mathcal{M}_g^{CA}$ 旋转平移至 $\mathcal{M}_{a\cap g}^{CA}$ 进行改善。

2.4 精细对齐

这里的精细对齐主要指的是从已经过粗略对齐的航拍与地面模型上获取可靠的三维对应点，并估计两模型之间的相似变换。考虑到通过基于图像建模获得的三维模型噪声较大，且模型会丢失二维图像所具有的丰富的纹理与上下文信息，本章采用通过二维图像匹配而非直接的三维模型匹配的方式实现航拍与地面模型的精细对齐。另外，考虑到原始航拍与地面图像之间在视角与尺度上的明显差异，直接对原始图像匹配进行效果并不理想。因此，本章采用另一种图像匹配方式以解决直接匹配原始图像导致的问题。具体来说，首先将地面模型投影至航拍图像视角下合成图像，然后对航拍图像与合成图像进行图像匹配，获取图像之间的二维匹配点。在此采用地面模型投影至航拍图像视角下以合成图像而非相反的原因是地面模型中的点云密度远大于航拍模型点云密度。另外，考虑到许多航拍图像覆盖了过于冗余的地面模型信息或者覆盖的信息与地面模型无关，因此，在进行二维图像匹配之前，应当将上述航拍图像进行剔除。因此，本章中基于二维图像匹配的三维模型精细对齐方法包含如下关键技术：

- 怎样选取合理覆盖地面模型的航拍图像子集；
- 怎样将地面模型投影至选取的航拍图像视角下以合成图像；
- 怎样获取合成图像与航拍图像之间的二维匹配点。

2.4.1 航拍图像选取

给出航拍图像集 $\{I_{a(i)}^{IN}|i = 1, 2, \ldots, N_a^{IN}\}$，如图2.3b 所示：一方面，由于航拍相机相对于地面相机通常分布区域更广，导致很多航拍图像并不覆盖地面模型

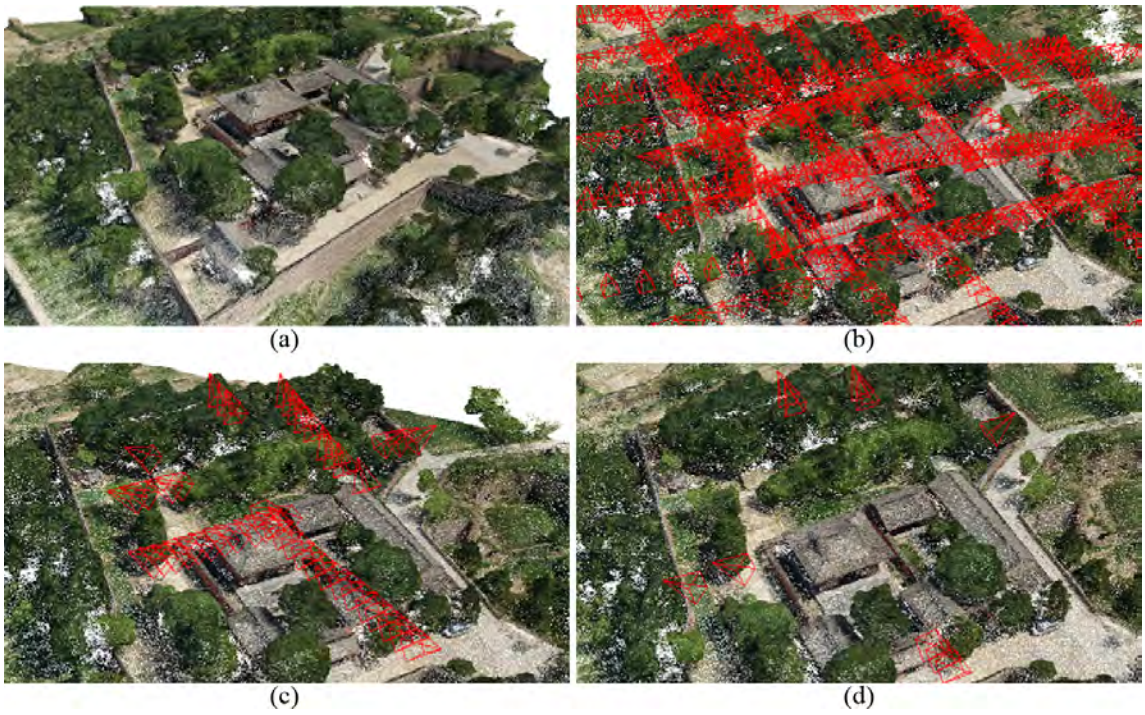

图 2.3: 航拍图像选取结果，其中红色棱锥表示相机位姿。(a) 航拍模型。(b) 采集的航拍图像集 $\{I^{IN}_{a(i)}|i=1,2,\ldots,N^{IN}_a\}$。(c) 初步选取的航拍图像集 $\{I^{IS}_{a(j)}|j=1,2,\ldots,N^{IS}_a\}$。(d) 最终选取的航拍图像集 $\{I^{FS}_{a(k)}|k=1,2,\ldots,N^{FS}_a\}$。

Figure 2.3: The result of aerial view selection, where the red cones denote the camera poses. (a) The aerial model. (b) The captured aerial image set $\{I^{IN}_{a(i)}|i=1,2,\ldots,N^{IN}_a\}$. (c) The initially selected aerial image set $\{I^{IS}_{a(j)}|j=1,2,\ldots,N^{IS}_a\}$. (d) The finally selected aerial image set $\{I^{FS}_{a(k)}|k=1,2,\ldots,N^{FS}_a\}$.

（图2.3a 中的寺庙）；另一方面，由于航拍摄影具有沿航线固定相机朝向进行拍摄的特性，许多航拍图像相似程度很高，导致这些图像包含了许多地面模型的冗余信息。因此，此处需要一种能够高效剔除上述两类航拍图像的图像选取方法以加速精细对齐过程。在选取航拍图像时，应考虑如下三个因素：

- 地面模型投影至航拍图像后得到的投影面积 $\{s^{CA}_{a(i)}|i=1,2,\ldots,N^{IN}_a\}$ 足够大以尽量完整地覆盖地面模型；
- 航拍相机的俯仰角 $\{\theta^{CA}_{a(i)}|i=1,2,\ldots,N^{IN}_a\}$ 应足够小以覆盖更多的地面模型立面信息；
- 航拍相机的位置 $\{\boldsymbol{c}^{CA}_{a(i)}|i=1,2,\ldots,N^{IN}_a\}$ 分布应足够均匀以在航拍与地面模型上获取分布均匀的三维空间对应点。

基于航拍与地面模型粗略对齐的结果以及上述三个因素，本章采用了一种自

动航拍图像选取方法。首先，考虑因素一和因素二，得到如图2.3c 所示的航拍图像初步选取结果。然后，通过考虑因素三获取如图2.3d 所示的最终选取的航拍图像子集。具体的图像选取过程如下所述。

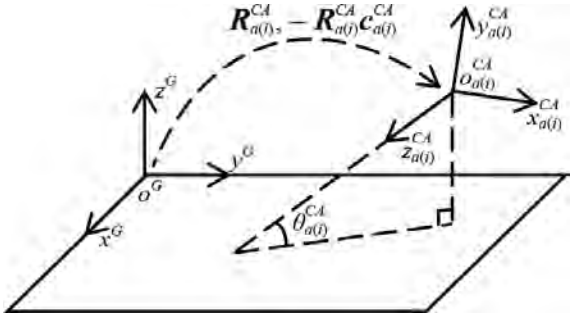

图 2.4: 俯仰角 $\theta^{CA}_{a(i)}$ 示意图。图中 $o^G - x^G y^G z^G$ 为地理坐标系，$o^{CA}_{a(i)} - x^{CA}_{a(i)} y^{CA}_{a(i)} z^{CA}_{a(i)}$ 为第 i 个经粗对齐的航拍相机坐标系。$[\boldsymbol{R}^{CA}_{a(i)} | -\boldsymbol{R}^{CA}_{a(i)} \boldsymbol{c}^{CA}_{a(i)}]$ 为上述两坐标系之间的欧式变换。
Figure 2.4: Sketch diagram of the pitch angle $\theta^{CA}_{a(i)}$. Here, $o^G = x^G y^G z^G$ is the geo-referenced coordinate system, and $o^{CA}_{a(i)} - x^{CA}_{a(i)} y^{CA}_{a(i)} z^{CA}_{a(i)}$ is the i-th coarsely aligned aerial camera coordinate system. $[\boldsymbol{R}^{CA}_{a(i)} | -\boldsymbol{R}^{CA}_{a(i)} \boldsymbol{c}^{CA}_{a(i)}]$ is the transformation between the above two coordinate systems.

对于航拍图像的初步选取，将经过粗略对齐的地面模型通过其包围盒近似表示。由于航拍与地面模型以及航拍相机均已经过粗略对齐，$s^{CA}_{a(i)}$ 可近似认为地面模型包围盒通过投影矩阵 $\boldsymbol{P}^{CA}_{a(i)} = \boldsymbol{K}^{CA}_{a(i)} [\boldsymbol{R}^{CA}_{a(i)} | -\boldsymbol{R}^{CA}_{a(i)} \boldsymbol{c}^{CA}_{a(i)}]$ 在航拍图像中得到的投影多边形面积。该投影面积与航拍图像总面积的比值记为 $r^{CA}_{a(i)}$。另外，航拍图像的俯仰角 $\theta^{CA}_{a(i)}$（见图2.4）可由航拍相机旋转矩阵 $\boldsymbol{R}^{CA}_{a(i)}$ 中抽取。

根据已获取的面积比 $r^{CA}_{a(i)}$ 及俯仰角 $\theta^{CA}_{a(i)}$，初步选取满足如下条件的航拍图像：

$$r^{CA}_{a(i)} > r_a, \theta^{CA}_{a(i)} < \theta_p, (i = 1, 2, \ldots, N^{IN}_a) \tag{2.3}$$

其中，面积比阈值 r_a 与俯仰角阈值 θ_p 在本章中分别设为 30% 与 45°。

假设经过初步航拍图像选取后，N^{IN}_a 幅航拍图像中有 N^{IS}_a 幅保留并将其记为 $\{I^{IS}_{a(j)} | j = 1, 2, \ldots, N^{IS}_a\}$。然后，通过考虑因素三，获取最终选取的航拍图像子集并记为 $\{I^{FS}_{a(k)} | k = 1, 2, \ldots, N^{FS}_a\}$。在获取最终航拍图像子集时，首先确定选取图像数目上限 N_s。如果 $N^{IS}_a \leq N_s$，本章方法将 $\{I^{FS}_{a(k)}\}$（最终选取航拍图像集合）置为 $\{I^{IS}_{a(j)}\}$；否则，将 $\{I^{FS}_{a(k)}\}$ 置为 $\{I^{FS}_{a(k)} = I^{IS}_{a(j^*_{(k)})} | k = 1, 2, \ldots, N_s\}$，即将初步选取的航拍图像集 $\{I^{IS}_{a(j)}\}$ 中的第 $j^*_{(k)}$ 幅图像选为最终选取的航拍图像集 $\{I^{FS}_{a(k)}\}$ 中的第 k 幅图像。最终选取的航拍图像集中的图像序号 $\{j^*_{(k)} | k = 1, 2, \ldots, N_s\}$ 通过如

下方式确定:

$$\begin{aligned} &\{j^*_{(k)}\} = \arg\max \sum_{m=1}^{3} \sigma_{(m)}(\omega_{(1)}\boldsymbol{c}^{IS}_{a(1)}, \omega_{(2)}\boldsymbol{c}^{IS}_{a(2)}, \ldots, \omega_{(N^{IS}_a)}\boldsymbol{c}^{IS}_{a(N^{IS}_a)}) \\ &s.t. \quad \omega_{(j)} = 0, 1; \quad \sum_{j=1}^{N^{IS}_a} \omega_{(j)} = N_s, (j = 1, 2, \ldots, N^{IS}_a) \end{aligned} \tag{2.4}$$

其中，$\sigma_{(m)}(\omega_{(1)}\boldsymbol{x}_{(1)}, \omega_{(2)}\boldsymbol{x}_{(2)}, \ldots, \omega_{(N)}\boldsymbol{x}_{(N)})$ 是加权点集 $\{\omega_{(j)}\boldsymbol{x}_{(j)}|j = 1, 2, \ldots, N\}$ 的协方差矩阵的第 m 个特征值，因此，$\sum_{m=1}^{3}\sigma_{(m)}$ 可用于度量点集的离散度。另外，$\omega_{(j)} = 0$ 表示点 $\boldsymbol{x}_{(j)}$ 不参与上述运算。因此，式2.4表示在 N^{IS}_a 幅航拍图像子集 $\{I^{IS}_{a(j)}\}$ 中选取 N_s 幅图像，使得选取的 N_s 幅图像相机位置离散度最大化。为兼顾计算效率以及最终选取航拍图像集合 $\{I^{FS}_{a(k)}\}$ 对地面模型覆盖的完整程度，N_s 在本章中设为 10。

算法 1: 航拍图像选取算法

Input :
采集的航拍图像集：$\{I^{IN}_{a(i)}|i = 1, 2, \ldots, N^{IN}_a\}$;
采集的航拍图像对应的相机参数：$\{\boldsymbol{K}^{IN}_{a(i)}, \boldsymbol{R}^{IN}_{a(i)}, \boldsymbol{c}^{IN}_{a(i)}|i = 1, 2, \ldots, N^{IN}_a\}$;
经粗略对齐的地面模型：$\mathcal{M}^{CA}_g$。
Output:
最终选取的航拍图像集：$\{I^{FS}_{a(k)}|k = 1, 2, \ldots, N^{FS}_a\}$;
最终选取的航拍图像的相机参数：$\{\boldsymbol{K}^{FS}_{a(k)}, \boldsymbol{R}^{FS}_{a(k)}, \boldsymbol{c}^{FS}_{a(k)}|i = 1, 2, \ldots, N^{IN}_a\}$。

根据式2.3获取初步选取的航拍图像集 $\{I^{IS}_{a(j)}|j = 1, 2, \ldots, N^{IS}_a\}$。
if $N^{IS}_a \leq N_s$ then
    $\{I^{FS}_{a(k)}\} \leftarrow \{I^{IS}_{a(j)}\}$，$N^{FS}_a \leftarrow N^{IS}_a$。
end
else
    根据式2.5设置 $\{I^{CS}_{a(l)}\}$ 中的第一幅图像，$N^{CS}_a \leftarrow 1$。
    repeat
        根据式2.6从 $\{I_{a(j)}{}^{IS}\}$ 选取一幅图像加入集合 $\{I^{CS}_{a(l)}\}$ 中，$N^{CS}_a \leftarrow N^{CS}_a + 1$。
    until $N^{CS}_a = N_s$;
    $\{I^{FS}_{a(k)}\} \leftarrow \{I^{IS}_{a(j)}\}$，$N^{FS}_a \leftarrow N^{IS}_a$。
end
return $\{I^{FS}_{a(k)}|k = 1, 2, \ldots, N^{FS}_a\}$，$\{\boldsymbol{K}^{FS}_{a(k)}, \boldsymbol{R}^{FS}_{a(k)}, \boldsymbol{c}^{FS}_{a(k)}|k = 1, 2, \ldots, N^{FS}_a\}$。

由于式2.4中定义的问题为一个 0-1 整数规划问题，该问题为 NP 问题，在这里本章方法采用一种贪婪算法对该问题进行近似求解。由于在航拍图像最终选取过程中，本章中的贪婪算法是对航拍图像进行逐一选取，为方便表述，在此定义一个动态图像集 $\{I^{CS}_{a(l)}|l = 1, 2, \ldots, N^{CS}_a\}$。这里，“动态集”指的是在最终航拍图像选取过程中，图像集合 $\{I^{IS}_{a(j)}\}$（共 N^{IS}_a 幅）中的 N^{CS}_a 幅被选中，加入到图像集合

$\{I^{CS}_{a(l)}\}$ 中，即图像集合 $\{I^{CS}_{a(l)}\}$ 中的图像数量 N^{CS}_a 是不断变化的。另外，在图像的贪婪选取过程中，集合 $\{I^{CS}_{a(l)}\}$ 中的第一幅图像，即集合 $\{I^{IS}_{a(j)}\}$ 中的第 $j^*_{(1)}$ 幅图像应最先确定。在此，本章方法通过如下方式进行确定：

$$j^*_{(1)} = \arg\max r^{IS}_{a(j)}/\theta^{IS}_{a(j)}, (j = 1, 2, \ldots, N^{IS}_a) \tag{2.5}$$

其中，$r^{IS}_{a(j)}$ 与 $\theta^{IS}_{a(j)}$ 分别为图像集合 $\{I^{IS}_{a(j)}\}$ 中第 j 幅图像的面积比与俯仰角。然后，本章方法通过如下方式，将图像集合 $\{I^{IS}_{a(j)}\}$ 中的第 $j^*_{N^{CS}_a+1}$ 幅图像选为图像集合 $\{I^{CS}_{a(l)}\}$ 中的第 $N^{CS}_a + 1$ 幅图像：

$$\begin{aligned} &j^*_{(N^{CS}_a+1)} = \arg\max \sum_{m=1}^{3} \sigma_{(m)}(\boldsymbol{c}^{CS}_{a(1)}, \boldsymbol{c}^{CS}_{a(2)}, \ldots, \boldsymbol{c}^{CS}_{a(N^{CS}_a)}, \boldsymbol{c}_{(j)}) \\ &s.t. \quad \boldsymbol{c}_{(j)} \in \complement_{\{\boldsymbol{c}^{IS}_{a(j)}\}}\{\boldsymbol{c}^{CS}_{a(l)}\}, (j = 1, 2, \ldots, N^{IS}_a - N^{CS}_a) \end{aligned} \tag{2.6}$$

其中，$\{\boldsymbol{c}^{IS}_{a(j)}\}$ 与 $\{\boldsymbol{c}^{CS}_{a(l)}\}$ 分别为在最终图像选取过程中初步选取与当前选取的图像对应的相机位置集合。$\complement_U A$ 表示相对于全集 U 的集合 A 的补集。式2.6中的过程不断重复，直至图像集合 $\{I^{CS}_{a(l)}\}$ 中的图像数量，即 N^{CS}_a 达到预设的数量上限 N_s。然后，本章方法将图像集合 $\{I^{FS}_{a(k)}\}$ 置为 $\{I^{CS}_{a(l)}\}$ 以完成最终航拍图像选取过程。另外，本章方法可同时获取最终选取的航拍图像的相机参数，记为 $\{\boldsymbol{K}^{FS}_{a(k)}, \boldsymbol{R}^{FS}_{a(k)}, \boldsymbol{c}^{FS}_{a(k)} | k = 1, 2, \ldots, N^{FS}_a\}$。算法1对本章中的航拍图像选取方法进行了概述。

2.4.2 航拍视角图像合成

给出最终选取的用于二维图像匹配的航拍图像集合 $\{I^{FS}_{a(i)} | i = 1, 2, \ldots, N^{FS}_a\}$（图2.7a 与图2.7e），可通过航拍视角图像合成方法生成对应的合成图像集合。这里的合成图像指的是在选取的航拍图像视角下的地面模型的投影。由于地面模型中的每个三维点都含有颜色信息，合成图像中每个像素的颜色信息即为投影到该像素的地面模型三维点的颜色信息。给出最终选取的航拍图像集合的投影矩阵 $\boldsymbol{P}^{FS}_{a(i)} = \boldsymbol{K}^{FS}_{a(i)}[\boldsymbol{R}^{FS}_{a(i)} | -\boldsymbol{R}^{FS}_{a(i)}\boldsymbol{c}^{FS}_{a(i)}]$，可获取经粗略对齐的地面模型 $\mathcal{M}^{CA}_g$ 中的每个空间三维点 $\boldsymbol{m}^{CA}_{g(j)}, (j = 1, 2, \ldots, N_g)$ 在选取的航拍图像视角下的投影点 $\boldsymbol{p}_{(i,j)}$：

$$\boldsymbol{p}_{(i,j)} \sim \boldsymbol{P}^{FS}_{a(i)}\boldsymbol{m}^{CA}_{g(j)}, (i = 1, 2, \ldots, N^{FS}_a; j = 1, 2, \ldots, N_g) \tag{2.7}$$

其中，$\sim$ 表示在相差一个尺度因子的意义下相等，N_g 为地面模型中的三维点个数。

由于航拍与地面模型已经过粗略对齐，航拍图像与对应的合成图像在透视投影形变，投影位置以及图像尺度方面差异均较小。另外，为实现模型的精细对齐，获取的二维图像匹配点需要转换成三维空间对应点，因此，还需要生成合成图像对应的深度图。

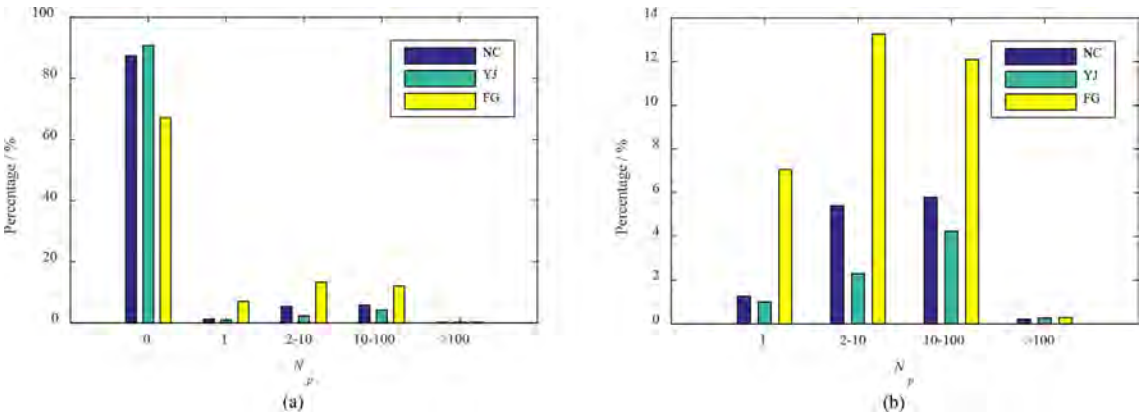

图 2.5: 地面模型在航拍图像视角下投影点的分布示例。其中，NC，YJ 与 FG 分别表示南禅寺，云居寺与佛光寺数据，在实验部分将对这些数据进行详细介绍。N_p 表示投影至单个航拍图像像素上的点的个数。（a）完整的投影点分布情况。（b）除去 $N_p = 0$ 情况的图（a）的放大图。

Figure 2.5: Projection distributions in the aerial view of the ground models. NC, YJ and FG are the datasets of Nan-Chan, Yun-Ju and Fo-Guang temples, respectively, which are detailed in the experimental section. N_p is the number of the projections onto one aerial image pixel. (a) The entire projection distribution. (b) The enlarged version of (a) without the case of $N_p = 0$.

然而，由于在基于图像的三维建模中噪声和外点是不可避免的，而这些噪声与外点会影响合成图像及对应的深度图进而使得图像匹配与模型对齐结果变差。另外，由于地面模型中的点云密度远大于航拍模型，因此在通过地面模型投影合成航拍视角图像时，常会出现多个（地面模型）空间点投影至同一（航拍图像）像素点的情况。为更直观的表述上述情况，图2.5给出了三个地面模型投影至航拍图像视角下投影点数量分布的示例，其中 N_p 表示投影至单个航拍图像像素上的点的个数。如图2.5a 所示，多数情况下 N_p 的值为 0，即没有地面模型中的点投影至这些航拍图像像素上，这是由于航拍图像覆盖的区域远大于地面模型。另外，由图2.5b 可知，相对于 $N_p = 1$，$N_p > 1$ 的情况更加常见。需要注意的是，N_p 的值有时甚至会大于 100。因此，多个地面模型三维点投影至同一航拍图像像素点的情况是普遍存在的。

基于上述分析，为使合成图像与对应的深度图效果更好，本章采用了一种同时考虑点深度及点密度的图像合成方法。在合成图像时考虑点密度的原因是密度大

的空间点更有可能位于物体表面而密度小的空间点则更有可能是噪声点或者外点。

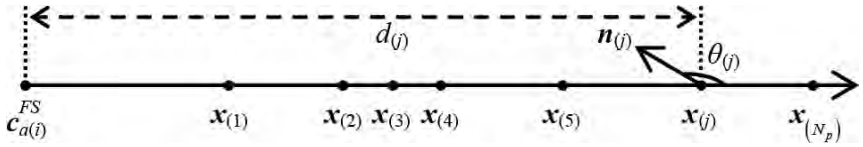

图 2.6: 点集 $\{\boldsymbol{x}_{(j)}|j=1,2,\ldots,N_p\}$ 投影至一幅选取的航拍图像 $I^{FS}_{a(i)}$ 中的一个像素的示意图。该图像光心为 $\boldsymbol{c}^{FS}_{a(i)}$。点集的深度、法向与可见角分别记为 $\{d_{(j)}|j=1,2,\ldots,N_p\}$，$\{\boldsymbol{n}_{(j)}|j=1,2,\ldots,N_p\}$ 与 $\{\theta_{(j)}|j=1,2,\ldots,N_p\}$。

Figure 2.6: Sketch diagram of the set of points $\{\boldsymbol{x}_{(j)}|j=1,2,\ldots,N_p\}$ with depths $\{d_{(j)}|j=1,2,\ldots,N_p\}$, normals $\{\boldsymbol{n}_{(j)}|j=1,2,\ldots,N_p\}$, and visible angles $\{\theta_{(j)}|j=1,2,\ldots,N_p\}$, that are projected onto a particular pixel of a selected aerial image $I^{FS}_{a(i)}$ with camera center $\boldsymbol{c}^{FS}_{a(i)}$.

如图2.6所示，点集 $\{\boldsymbol{x}_{(j)}|j=1,2,\ldots,N_p\}$（其深度、法向与可见角分别为 $\{d_{(j)}|j=1,2,\ldots,N_p\}$，$\{\boldsymbol{n}_{(j)}|j=1,2,\ldots,N_p\}$ 与 $\{\theta_{(j)}|j=1,2,\ldots,N_p\}$）为 N_p 个投影至待合成图像的同一像素的空间点。点的深度为点 $\boldsymbol{x}_{(j)}$ 到相机光心 $\boldsymbol{c}^{FS}_{a(i)}$ 的距离；点的可见角为由 $\boldsymbol{c}^{FS}_{a(i)}$ 指向 $\boldsymbol{x}_{(j)}$ 的射线与 $\boldsymbol{x}_{(j)}$ 的法向 $\boldsymbol{n}_{(j)}$ 的夹角。空间点相对于给定的相机的可见性可通过可见角衡量，即当点 $\boldsymbol{x}_{(j)}$ 的可见角 $\theta_{a(j)}>\theta_v$ 时，其在选取的航拍图像 $I^{FS}_{a(i)}$ 是可见的。θ_v 在本章中的值设为 90°。然后，本章方法通过如下的优化过程选取投影至当前像素的序号为 j^* 的空间点用作图像合成与深度图构建：

$$j^*=\arg\max(t_{d(j)}+\alpha t_{\rho(j)}),(j=1,2,\ldots,N_p) \tag{2.8}$$

其中，α 为加权系数，本章中将其值设置为 0.5。$t_{d(j)}$ 与 $t_{\rho(j)}$ 分别为点 $\boldsymbol{x}_{(j)}$ 的归一化深度项与密度项，定义如下：

$$t_{d(j)}=\frac{d_{(N_p)}-d_{(j)}}{d_{(N_p)}-d_{(1)}},t_{\rho(j)}=\frac{\rho_{(j)}-\min\{\rho_{(j)}\}}{\max\{\rho_{(j)}\}-\min\{\rho_{(j)}\}},(j=1,2,\ldots,N_p) \tag{2.9}$$

由上述定义可知，归一化深度项 $t_{d(j)}$ 与该点的深度值 $d_{(j)}$ 负相关而归一化密度项 $t_{\rho(j)}$ 与该点的密度值 $\rho_{(j)}$ 正相关。因此，经过式2.8定义的优化过程，有着更小的深度值 $d_{(j)}$ 与更大密度值 $\rho_{(j)}$ 的点会被优先选取。式2.9中的点的密度值 $\rho_{(j)}$ 定义如下：

$$\rho_{(j)}=\begin{cases}\frac{1}{d_{(j+1)}-d_{(j)}},j=1\\ \frac{1}{d_{(j)}-d_{(j-1)}},j=N_p\\ \frac{2}{d_{(j+1)}-d_{(j-1)}},\text{otherwise}\end{cases} \tag{2.10}$$

由式2.10可知，点 $\boldsymbol{x}_{(j)}$ 的密度值 $\rho_{(j)}$ 取决于该点临近点的深度差。

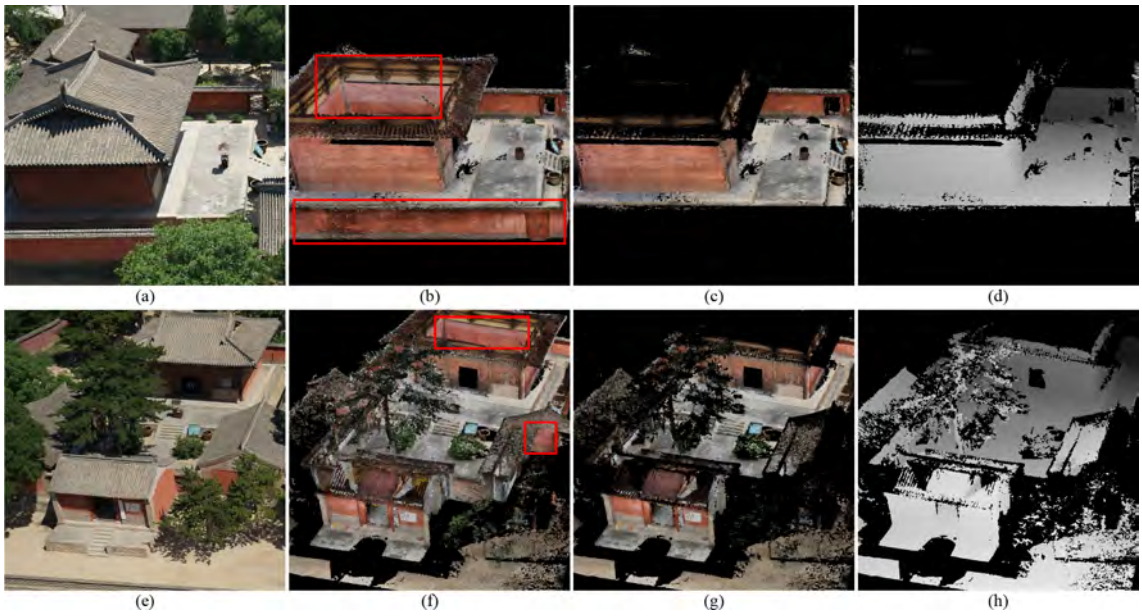

图 2.7: 航拍图像合成结果。（a）采集的航拍图像。（b）未经可见性滤波的合成图像。（c）经可见性滤波的合成图像。（d）构建的深度图。（e）-（h）另一个航拍图像合成结果示例。图（b）与图（f）中的红色矩形标示出了合成图像的错误区域。

Figure 2.7: The aerial view synthesis result. (a) The captured aerial image. (b) The synthesized image without visibility filtering. (c) The synthetic image with visibility filtering. (d) The constructed depth map. (e)-(h) Another example of aerial view synthesis result. The red rectangles in (b) and (f) highlight the artifacts of image synthesis.

另外，对于上述航拍视角图像合成方法来说，还存在一个问题。由于本章方法将地面模型中的点投影至航拍视角以合成图像，因此有些对于给定航拍相机来说处于地面模型背面的点也投影到了当前航拍视角下，这会导致错误的图像合成结果，见图2.7b 与图2.7f 中的红色矩形。

为解决上述问题，当合成航拍视角图像时，本章方法只保留对于当前航拍视角可见点对应的像素。另外，经过上述图像合成与滤波过程，可以建立三维空间点与二维图像点之间的对应关系并顺便将合成图像对应的深度图构建出来。图2.7c，图2.7d 与图2.7g，图2.7f 为两个经过可见性过滤的合成图像与对应深度图示例。如图所示，上述可见性滤波过程消除了绝大多数图像合成时产生的错误。尽管由于法向估计噪声或者点云的空间不连续特性导致合成图像与对应的深度图仍有一些不可避免的错误区域，这些区域一般都较小且对后续的图像匹配过程几乎没有影响。在图像合成与深度图构建之后，本章方法对其进行中值滤波，以减少其噪声并对可能出现的图像孔洞进行填补。

2.4.3 子区域 SIFT 匹配

给出合成的航拍视角图像，可对合成图像与航拍图像进行图像匹配。由于本章方法通过将地面模型投影至航拍视角下的方式进行图像合成，而航拍图像覆盖的区域远大于地面模型。因此，合成的航拍视角图像中有很多无效区域。在此，本章方法只对合成图像与航拍图像的公共区域进行图像匹配。此公共区域设为地面模型包围盒在航拍图像中的投影区域。

然而，由于公共区域面积较大且合成图像噪声较大，直接对整个公共区域进行图像匹配得到的结果中正确匹配点占的比例较低。由于航拍与地面模型已经过粗略对齐，同一物体在合成图像与航拍图像中的投影位置不会偏差过大，因此，在此采用子区域逐一匹配而非全区域一次性匹配的方式，采用 SIFT 特征，实现图像匹配。在进行图像匹配时，通过 FLANN[56] 寻找最近邻点，且对得到的匹配点进行交叉验证及 NNDR 验证 [76]。其中，在本章中比值阈值 r_n 设为 0.8。对于分辨率为 $height \times width$ 的待匹配图像区域，此处对大小为 $height/N_r \times width/N_r$ 的子区域逐对进行 SIFT 匹配操作。本章中将 N_r 的值设为 10。

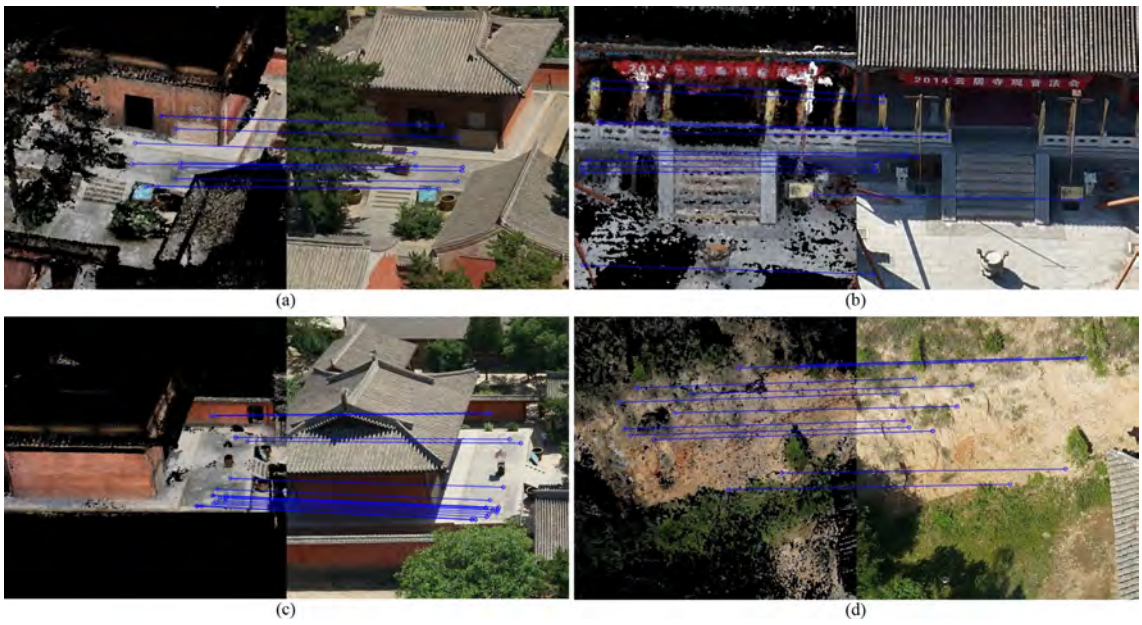

图 2.8: 四对本章图像匹配方法结果示例。其中，蓝色线段用来标示匹配点。
Figure 2.8: Four sample image matching results of the proposed method. The blue segments denote the putative image matches.

另外，由于合成图像与选取的航拍图像之间的透视投影形变及尺度差异均较

小，本章进一步对候选匹配点进行过滤，不满足如下条件的匹配点将被滤除：

$$\sigma_q < \sigma_{s(i)}/\sigma_{a(i)} < 1/\sigma_q, |\theta_{s(i)} - \theta_{a(i)}| < \theta_d, (i = 1, 2, \ldots, N_m) \tag{2.11}$$

其中，$\{\sigma_{s(i)}|i = 1, 2, \ldots, N_m\}$ 与 $\{\sigma_{a(i)}|i = 1, 2, \ldots, N_m\}$ 分别是合成图像与航拍图像中匹配点的尺度；$\{\theta_{s(i)}|i = 1, 2, \ldots, N_m\}$ 与$\{\theta_{a(i)}|i = 1, 2, \ldots, N_m\}$ 是匹配点的主方向。另外，尺度比值阈值 σ_q 与主方向偏差阈值 θ_d 在本章中分别设为 0.8 与 30°。图2.8给出了四对图像匹配结果示例，其中，蓝色线段用来标示匹配点。由图可知，通过本章的图像匹配方法，可获取足够多的合成图像与航拍图像的匹配点。

在获取合成图像与航拍图像之间的二维匹配点之后，经粗略对齐的航拍模型与地面模型之间的三维对应点可通过图像对应的深度图获取。之后，可通过 RANSAC 估计相似变换实现航拍与地面模型的精细对齐。在进行 RANSAC 模型估计时，迭代次数与内点阈值 ε^{FA} 分别设为 500 与 $0.3m$。最后，将经粗略对齐的地面模型根据估计得到的相似变换变换至航拍模型坐标系下，实现模型的精细对齐:

$$\mathcal{M}_g^{FA} = \mathbb{S}_g^{FA}\mathbb{R}_g^{FA}\mathcal{M}_g^{CA} + \mathbb{T}_g^{FA} \tag{2.12}$$

其中，$\mathcal{M}_g^{FA}$ 为经精细对齐的地面模型；而航拍模型维持经粗略对齐后的状态，即 $\mathcal{M}_a^{FA} = \mathcal{M}_a^{CA}$；$\{\mathbb{S}_g^{FA}, \mathbb{R}_g^{FA}, \mathbb{T}_g^{FA}\}$ 为用于模型精细对齐的相似变换。

2.5 实验结果

2.5.1 实验数据

正如文献 [33] 中所述，目前几乎没有覆盖同一区域的航拍与地面图像公开数据集。因此，本章中在自建数据集上进行方法评测。数据集为三个中国古代寺庙：南禅寺（NC，图2.9a 与图2.9b），云居寺（YJ，图2.9f 与图2.9g）与佛光寺（FG，图2.9k 与图2.9l）。三组图像数据集的具体细节列于表2.2。由表2.2可知，航拍与地面模型在视角（俯仰角）与尺度（空间分辨率）方面差异巨大。另外，本章中的航拍与地面模型是采用方法 [73, 75] 进行重建得到的。

表 2.2: 用于方法测评的三组图像数据集的具体细节。其中，$a:b:c$ 表示以 a 为起点，c 为终点，b 为步长的一组数。

Table 2.2: Details of the three image collections for evaluating the proposed method. $a:b:c$ denotes a set of numbers, whose beginning is a, ending is c, and step is b.

数据集	南禅寺（NC）	云居寺（YJ）	佛光寺（FG）
航拍图像数量	1429	1347	1596
地面图像数量	2790	936	972
航拍图像采集模式	5 条航线：1 条用于采集垂直视角图像，4 条用于采集 45° 倾斜视角图像 每站 45 幅图像	每站 12 幅图像	每站 36 幅图像
地面图像采集模式	俯仰角：$-40°:20°:40°$ 偏航角：$0°:40°:320°$	俯仰角：$0°$ 偏航角：$0°:30°:330°$	俯仰角：$-20°:20°:40°$ 偏航角：$0°:40°:320°$
航拍/地面图像采集设备	安装在 UAV 上的 Sony NEX-5R/Canon EOS 5D Mark III		
航拍/地面相机焦距	$24mm/35mm$		
航拍/地面图像分辨率	$4912px\times3264px/5760px\times3840px$		
航拍/地面图像地理对齐方式	差分 GPS 测量的 GCP 地理坐标/相机内置 GPS		
航拍图像空间分辨率	$8.47mm/px$	$13.16mm/px$	$14.29mm/px$
地面图像空间分辨率	$0.73mm/px$	$0.91mm/px$	$0.77mm/px$

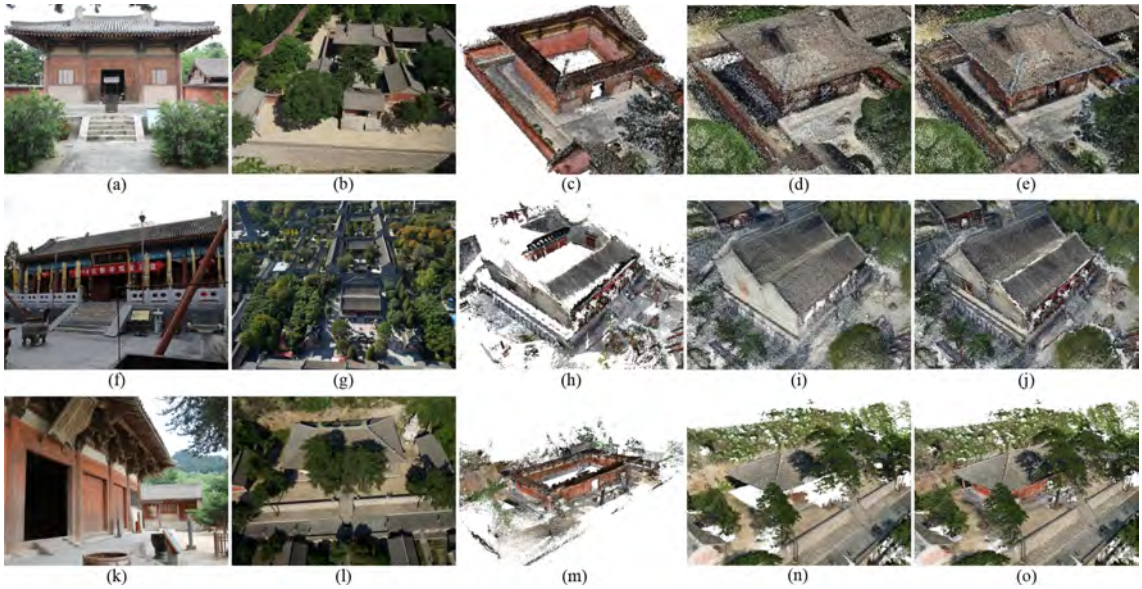

图 2.9: 用于评测的航拍与地面图像集与本章的航拍与地面模型精细对齐方法结果。(a) 南禅寺的一张地面示例图像。(b) 南禅寺的一张航拍示例图像。(c) 南禅寺地面模型。(d) 南禅寺航拍模型。(e) 南禅寺精细对齐结果。(f) - (j) 云居寺的类似 (a) - (e) 的图例。(k) - (o) 佛光寺的类似 (a) - (e) 的图例。

Figure 2.9: Image collections for evaluation and the fine alignment results of the proposed method in this chapter. (a) A sample ground image of Nan-Chan temple. (b) A sample aerial image of Nan-Chan temple. (c) Ground model of Nan-Chan temple. (d) Aerial model of Nan-Chan temple. (e) Fine alignment result of Nan-Chan temple. (f)-(j) Items similar to (a)-(e) of Yun-Ju temple. (k)-(o) Items similar to (a)-(e) of Fo-Guang temple.

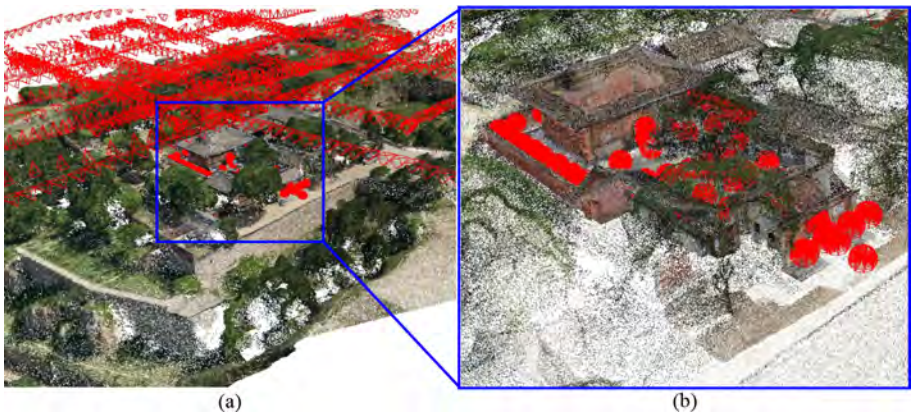

图 2.10: 南禅寺的模型与相机（红色棱锥）精细对齐结果。（a）远景。（b）近景。
Figure 2.10: The fine alignment results of the models and camera poses (red cones) for Nan-Chan temple. (a) Zoom-out result. (b) Zoom-in result.

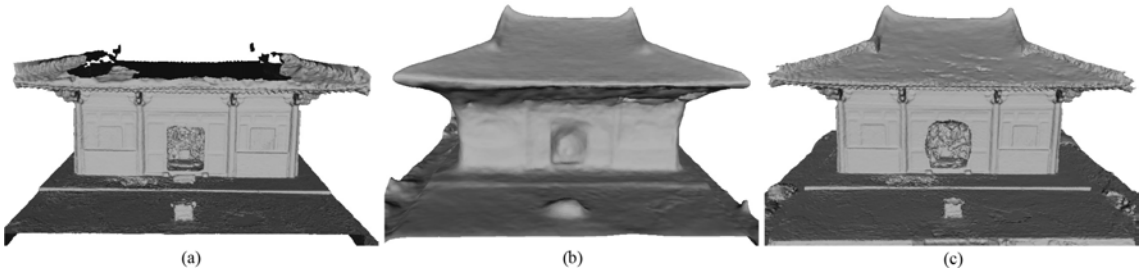

图 2.11: 南禅寺数据表面重建结果。（a）地面模型。（b）航拍模型。（c）经精细对齐后的航拍与地面模型。
Figure 2.11: Surface reconstruction results for the Nan-Chan temple. (a) The ground model. (b) The aerial model. (c) The ground and aerial models after fine alignment.

2.5.2 模型对齐定性结果

南禅寺，云居寺与佛光寺精细对齐的定性结果分别如图2.9e，2.9j 与2.9o 所示。另外，图2.10给出了南禅寺的模型与相机精细对齐结果。由图2.9与图2.10可知，航拍与地面的模型与相机均较好的对齐到了一起。为体现本章航拍与地面模型对齐方法的有效性，此处采用 [7] 中的方法对对齐后的南禅寺模型进行了表面重建，如图2.11所示。其中，图2.11a，图2.11b 与图2.11c 分别为对南禅寺地面模型，航拍模型与经精细对齐后的航拍与地面模型进行表面重建的结果。由图可见，地面模型的表面重建结果缺少屋顶信息而航拍模型的表面重建结果缺少建筑物立面结构细节。而经精细对齐的航拍与地面模型的重建结果在完整性与细节方面均取得了较好的结果。

图 2.12: 用于定量评价的参照点示例。（a）南禅寺数据的参照点示例。（b）佛光寺数据的参照点示例。
Figure 2.12: Examples of reference points for quantitative evaluation. (a) Reference points of Nan-Chan temple. (b) Reference points of Fo-Guang temple.

2.5.3 定量评价指标

2.5.4 参数设定评测

表 2.3: 本章方的参数表。
Table 2.3: Parameters table of the proposed method in this chapter.

符号	取值	章节	描述
ε^{CA}	$1.0m$	2.4	用于估计进行粗略对齐的相似变换的距离阈值
r_a	30%	2.4.1	用于初步航拍图像选取的面积比阈值
θ_p	45°	2.4.1	用于初步航拍图像选取的俯仰角阈值
N_s	10	2.4.1	最终选取的航拍图像数量上限
α	0.5	2.4.2	用于图像合成的加权系数
θ_v	90°	2.4.2	用于可见性过滤的可见角阈值
r_n	0.8	2.4.3	用于 NNDR 验证的比值阈值
N_r	10	2.4.3	用于图像匹配的单维度子区域个数
σ_q	0.8	2.4.3	用于过滤候选匹配点的尺度比值阈值
θ_d	10	2.4.3	用于过滤候选匹配点的主方向偏差阈值
ε^{FA}	$0.3m$	2.4.3	用于估计进行精细对齐的相似变换的距离阈值

正如文献 [4] 中提到的，对模型对齐精度定义一个确切的定量评价指标是十分困难的。本节设计了一个用于定量评价对齐精度的近似评价指标。首先，人工在航拍与地面模型中选取一些对应点并将它们标注为参照点。选取的参照点应当相对均匀地分布在整个航拍与地面模型的重叠区域上。图2.12给出了南禅寺数据与佛光寺数据的一些参照点示例，由红色叉号标示。然后，将地面模型上的参照点根据估计得到的，用于模型对齐的相似变换进行变换。最后，求取变换后的地面模型参照

点与对应的航拍模型参照点之间的距离。此处将该组距离的均值与中值用作定量评价指标。值得注意的是，上述评价指标不仅包含航拍与地面模型对齐的误差，也包含了由于重建算法和人工操作导致的误差。但是，该评价指标在通常情况下足以衡量模型对齐精度。

本章算法共涉及到 11 个参数，如表2.3所示。它们中的 5 个参数（N_s，α，N_r，σ_q 与 θ_d）对模型对齐结果影响相对较大。接下来，本节将会对这 5 个参数的设定进行评测。其它的参数对结果影响相对较小，因此，在模型对齐过程中，本节将它们的值固定为表2.3中的值。

2.4.1节中的 N_s，2.4.2节中的 α 与2.4.3节中的 N_r 直接影响图像选取，图像合成与图像匹配的结果。并且，所有上述中间步骤都会对最终的模型对齐结果产生较大的影响。因此，此处通过2.5.3节中介绍的模型对齐均值与中值误差来评测上述三个参数的设置。由于 N_s，α 与 N_r 三个参数分别属于本章方法中的三个连续子步骤，在此对它们进行逐一测评。在测评某一参数时，其它两个参数的值固定且设为表2.3中的值。上述三个参数的测评结果如图2.13所示。

由图2.13a 与图2.13b 可知，随着 N_s 的增大，对齐精度逐渐增加，但计算时间也会相应增加。当 $N_s > 10$ 时，对齐精度的变化不再明显。因此，为平衡精度与效率，本章中 N_s 的值设置为 10。

参数 α 的引入，即式2.8中归一化密度项 $t_{\rho(j)}$ 的引入，对对齐精度的提升相当明显。如图2.13a 与图2.13b 所示，当 $\alpha > 0.2$ 时，α 值的变化对对齐精度影响较小，将 α 的值设为 0.5 时可以取得较好的对齐精度。

由图2.13e 与图2.13f 可知，参数 N_r 的值也会对对齐精度产生较大的影响。当 N_r 的值过小或过大时，对齐精度均较差。一方面，随着 N_r 减小，用于 SIFT 匹配的子区域变大，而在相对较大的区域上对选取图像与噪声较大的合成图像进行图像匹配会导致相对较差的匹配结果。当 $N_r = 1$ 即在整个区域进行 SIFT 匹配，因此对齐精度很差。另一方面，随着 N_r 增大，子区域相应变小。当 N_r 过小时，由于粗对齐不够精确，选取图像与合成图像之间的用于 SIFT 匹配的子区域重叠区域会变得过小。当 N_r 的值设置在 7 至 17 之间时，对齐精度较好，而在本章中 N_r 的值设为 10。

另外，式2.11给出的过滤条件会影响 SIFT 匹配，进而影响模型对齐的结果。在此通过对两个过滤条件逐一禁用并将模型对齐结果与原始结果进行比较的方式测评上述过滤条件。对比结果如表2.4所示。由表2.4可知，本章的匹配点过滤方法

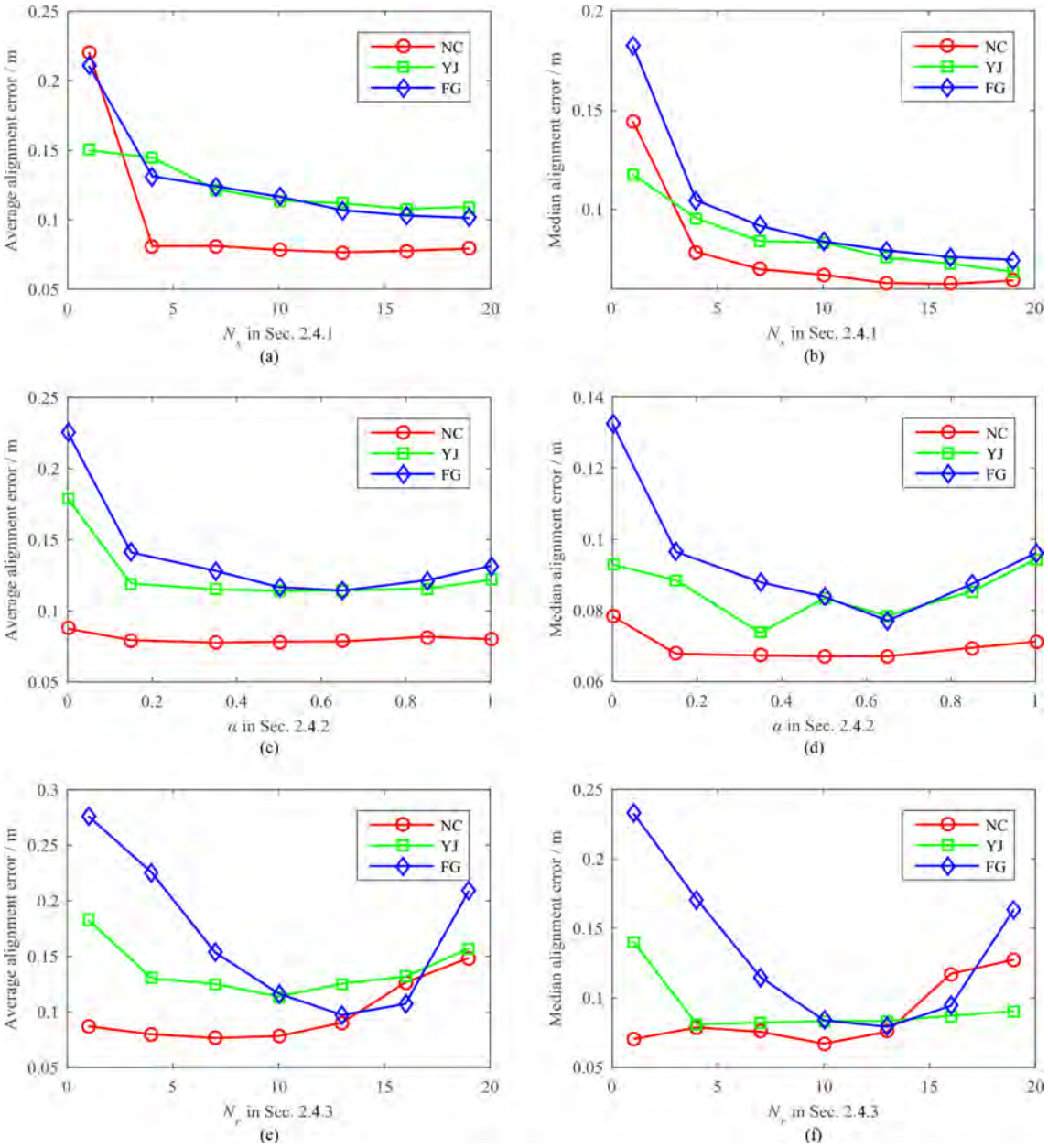

图 2.13: 参数设定评测实验结果。测评的参数包括2.4.1节中的 N_s，2.4.2节中的 α 与2.4.3节中的 N_r。图中的 NC，YJ 与 FG 分别指的是南禅寺，云居寺与佛光寺数据集。左边一列与右边一列分别表示以米为单位的模型对齐均值误差与中值误差。

Figure 2.13: The experimental results for parameter setting evaluation. The parameters evaluated here are N_s in Section 2.4.1, α in Section 2.4.2 and N_r in Section 2.4.3. NC, YJ and FG are the datasets of Nan-Chan, Yun-Ju and Fo-Guang temple. The left and right columns show the average and median alignment errors (in meters), respectively.

可提升 SIFT 匹配结果，进而提升模型对齐精度。

表 2.4: 式2.11中 σ_q 与 θ_d 的评测结果。其中，$\bar{x}$ 与 $\tilde{x}$ 分别表示以米为单位的模型对齐均值误差与中值误差。

Table 2.4: Evaluation of parameters σ_q and θ_d in Eq. 2.11. $\bar{x}$ and $\tilde{x}$ are the average and median alignment errors (in meters) respectively.

数据集	本章方法		禁用式2.11中的 σ_q		禁用式2.11中的 θ_d	
	$\bar{x}/m$	$\tilde{x}/m$	$\bar{x}/m$	$\tilde{x}/m$	$\bar{x}/m$	$\tilde{x}/m$
NC	**0.0783**	**0.0671**	0.0966	0.0898	0.0866	0.0796
YJ	**0.1139**	**0.0834**	0.1311	0.0992	0.1261	0.0957
FG	**0.1166**	**0.0839**	0.1321	0.1251	0.1211	0.1063

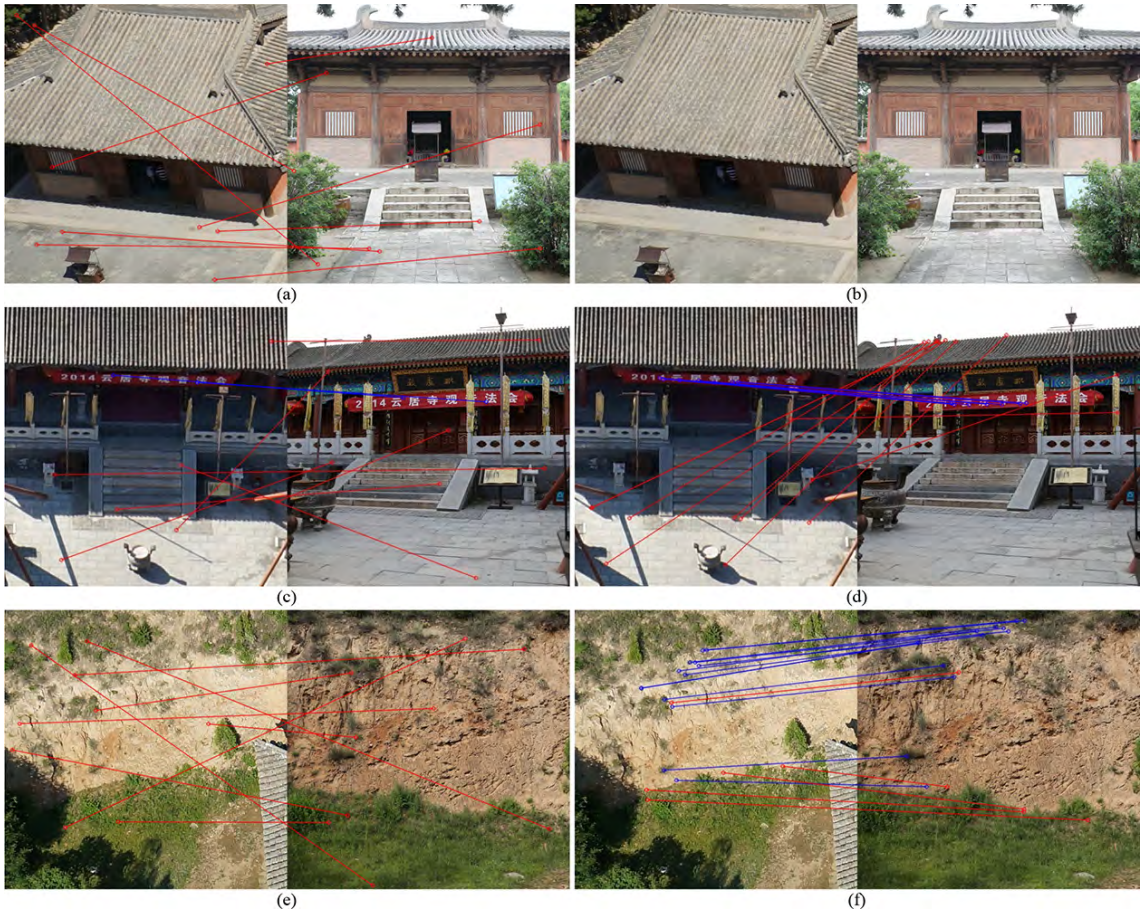

图 2.14: 三对航拍与地面图像对之间的图像匹配结果。(a)，(c) 与 (e) SIFT[1] 图像匹配结果。(b)，(d) 与 (f) ASIFT[2] 图像匹配结果。其中，蓝色线段表示正确匹配点而红色线段表示错误匹配点。

Figure 2.14: The image matching results between three ground and aerial image pairs. (a), (c), and (e) The results of SIFT[1] matching. (b), (d), and (f) The results of ASIFT[2] matching. The blue segments denote the true-positive point matches while the red segments denote the false-positive point matches.

2.5.5 与现有方法比较

本节对本章的方法与以下三类方法进行比较：（1）基于二维局部特征的图像匹配方法 [1, 2]；（2）基于三维局部特征的模型对齐方法 [3] 以及（3）基于三维模型合成二维图像的方法 [4]。

对于基于二维局部特征的图像匹配方法，本节尝试采用 SIFT 特征 [1] 与 ASIFT 特征 [2] 直接匹配航拍与地面图像。然而，由于航拍与地面图像之间的视角与尺度差异过大，在原始图像上进行图像匹配几乎不能获得正确的匹配点。因此，为消除尺度差异这个因素的影响，本节首先将航拍图像剪裁至只覆盖与地面图像相同区域并将地面图像采样至与剪裁后的航拍图像接近的分辨率，然后对经过上述预处理的航拍与地面图像进行 SIFT 与 ASIFT 特征匹配以及后续的 NNDR 验证及基本矩阵过滤。图2.14给出了通过上述方法得到三对图像匹配结果示例。由图2.14可知，尽管经过预处理，在进行航拍与地面图像匹配时，二维局部特征点的表现仍然较差。由图2.14左边一列可知，SIFT 图像匹配结果十分混乱，这是由于 SIFT 特征不能够应对较大的视角变化。而由图2.14右边一列可知，ASIFT 图像可以获得一些正确的匹配点，这是由于相对于 SIFT 特征，ASFIT 特征对视角变化更为鲁棒。但是，即使如此，相对于本章中的图像匹配方法（见图2.8），基于 ASIFT 的图像匹配结果仍不能让人满意。

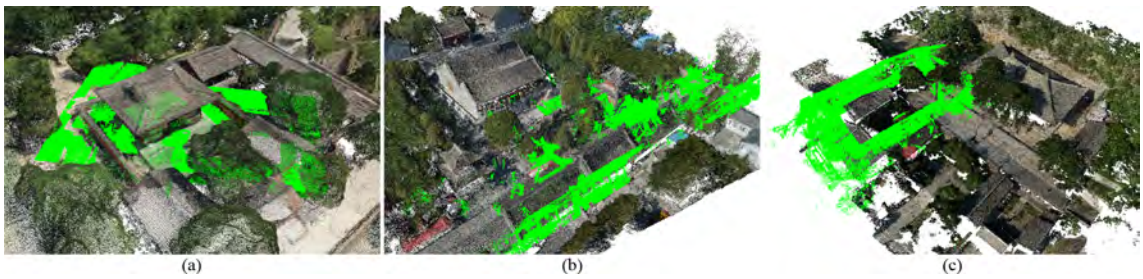

图 2.15: 基于 FPFH 特征 [3] 的模型对齐结果。（a）南禅寺数据结果。（b）云居寺数据结果。（c）佛光寺数据结果。为了更好的视觉效果，图中的地面模型显示为绿色。
Figure 2.15: The results of the FPFH[3] based model alignment method. (a) The result for Nan-Chan temple. (b) The result for Yun-Ju temple. (c) The result for Fo-Guang temple. The ground models are colored green for better visualization. The alignment results are completely incorrect.

对于基于三维局部特征的模型对齐方法，本节采用的是 FPFH 特征 [3]。该方法的输入为待对齐的航拍与地面模型。本节首先对点云降采样，然后在降采样的点云上提取 FPFH 特征，进而通过基于 RANSAC 的相似变换估计的方式将两模型进行对齐。本节采用上述方法在南禅寺，云居寺与佛光寺数据上进行模型对齐，

结果如图2.15所示。由图2.15可知，上述方法的对齐结果是完全错误的。导致这种情况的原因概括如下：（1）航拍与地面模型的重叠区域非常有限；（2）地面模型中的点云密度远大于航拍模型；（3）通过基于图像建模的流程得到的模型通常噪声较大。所有上述的因素导致三维局部特征用于航拍与地面模型对齐的效果较差。

表 2.5: 本章方法与方法 [4] 的模型对齐结果对比。其中，$\bar{x}$ 与 $\tilde{x}$ 分别表示以米为单位的模型对齐均值误差与中值误差；T 表示以秒为单位的总的计算时间。
Table 2.5: The comparison results for model alignment with Shan et al. [4]. $\bar{x}$ and $\tilde{x}$ are the average and median alignment errors (in meters). T is the total time-cost (in seconds).

数据集	本章方法			Shan et al.		
	$\bar{x}/m$	$\tilde{x}/m$	T/s	$\bar{x}/m$	$\tilde{x}/m$	T/s
NC	**0.0783**	**0.0671**	**426**	0.0919	0.0907	1715
YJ	**0.1139**	**0.0834**	**441**	0.1331	0.1158	1754
FG	**0.1166**	**0.0839**	**455**	0.1287	0.1075	1877

最后，本节将本章方法与一种原理类似的，基于三维模型合成二维图像的方法 [4] 在精度与效率上进行了对比。实验结果如表2.5所示。由表2.5可知，本章方法在对齐精度方面高于 [4]。这是由于在本章中，航拍视角图像是通过地面模型投影合成，合成的图像更加完整且更适用于进行 SIFT 图像匹配。至于对齐效率，相对于本章方法，方法 [4] 更加耗时（将近慢 4 倍）。这是由于方法 [4] 的时间复杂度为 $o(N_s N_s')$ 而本章方法的时间复杂度为 $o(N_s)$，其中，N_s 月 N_s' 分别为在模型对齐过程中涉及到的航拍与地面图像数量。

2.6 本章小结

本章提出了一种基于稠密点云的航拍与地面模型精确、高效的对齐方法。该方法包含两个步骤：粗略对齐与精细对齐。粗略对齐通过利用航拍与地面图像的 GPS 信息将航拍与地面模型变换至地理坐标系实现。精细对齐根据经粗略对齐的模型上的三维对应点估计相似变换实现。本章的主要贡献为在精细对齐过程中，寻找可靠的三维对应点时的三个关键步骤，即航拍图像选取，航拍视角图像合成以及子区域 SIFT 匹配。实验结果表明本章方法可有效地实现航拍与地面模型的对齐，且相比于其他方法，本方法在对齐精度与效率方面均表现更好。

第 3 章　基于稀疏点云的航拍与地面点云融合

尽管上一章中的基于稠密点云的航拍与地面模型对齐方法可以实现航拍与地面模型的精确、高效对齐，该方法仍存在一些问题：（1）采用较为耗时的 MVS 获取地面稠密点云；（2）通过稠密点云投影进行航拍视角图像合成时会存在像素缺失现象；（3）通过估计相似变换实现模型对齐，难以应对基于图像重建中存在的场景漂移现象。针对上述问题，本章提出了一种基于稀疏点云的航拍与地面点云融合方法。

3.1 引言

基于图像的大规模建筑场景重建在计算机视觉 [5, 6, 72, 77] 与遥感 [78–80] 领域是一个十分经典与基础的问题。由于近年来在算法效率与硬件性能上的迅猛发展，当前的重建系统已经由重建单体建筑扩展为重建整个城市 [81]，甚至全世界 [82]。为了兼顾重建模型的完整性与细节程度，通常的做法是分别利用航拍与地面图像进行大尺度与近距离的图像采集并重建得到航拍与地面点云，然后将两种点云进行融合。考虑到通过基于图像建模获得的三维模型噪声较大的特性，且三维模型会丢失二维图像所具有的丰富的纹理与上下文信息，通过二维图像匹配的方式进行点云融合比直接三维点云对齐的方式（例如 ICP[30]）更为可取。

图3.1展示了融合航拍与地面点云的难点，其中图3.1a 为一对航拍与地面图像示例，图3.1b 为由图像重建得到的航拍与地面稀疏点云。这里的稀疏点云指的是通过 SfM 对图像局部特征点（例如 SIFT）重建得到的点云。图3.1展示了采用航拍与地面图像进行场景重建的两个关键问题：

- 如何匹配在视角和尺度上差异巨大的航拍与地面图像（图3.1a）；
- 如何融合在噪声程度、密度和精度方面有明显差异的航拍与地面点云（图3.1b）。

为解决航拍与地面图像匹配问题，本章中采用的方案仍然为将地面图像变换至航拍视角下以消除两类图像在视角与尺度方面的差异。本章方法与方法 [4] 以及第二章的方法有本质不同。上述两种方法的航拍视角图像合成基于空间离散的地面稠密点云，而本章的方法基于空间连续的地面稀疏网格，该网格由地面稀疏点

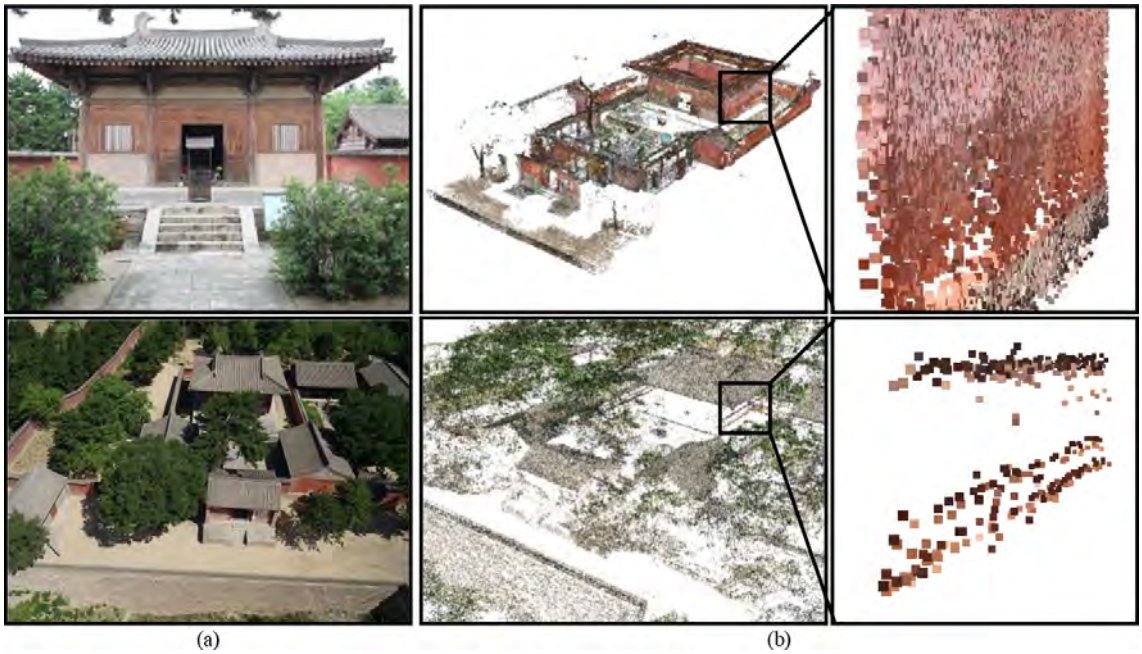

图 3.1: 南禅寺示例航拍与地面图像以及对应的稀疏点云。（a）示例航拍与地面图像。（b）航拍与地面稀疏点云。图（b）中右边一列为图（b）中左边一列黑色矩形区域的放大图。
Figure 3.1: Ground and aerial images and sparse point clouds of Nanchan Temple. (a) Example ground and aerial images. (b) Ground and aerial sparse point clouds. The images in right column of (b) are the enlarged patches in the black rectangles of the images in the left column of (b).

云进行表面重建得到。这样做的好处是可以避免稠密点云计算的耗时操作。之后，本章方法采用 SIFT 特征匹配航拍图像与合成图像。最后，本章方法通过如下两步对候选匹配点进行过滤：（1）匹配点之间的尺度与主方向一致性检验；（2）匹配点位置之间的仿射变换验证。

在匹配完航拍与地面图像之后，本章方法通过全局 BA，而非估计相似变换，实现航拍与地面点云的融合。这样做的好处是可以在一定程度上应对场景飘移问题且能够取得更加精确的融合结果。为实现上述目的，本章方法首先将得到的航拍与地面图像匹配点连入原始航拍特征点轨迹，然后对增广航拍特征点轨迹与原始地面特征轨迹进行全局 BA，实现航拍与地面点云的融合。

本章方法的主要由如下三个主要贡献：

- 本章中的航拍视角图像基于地面稀疏网格进行合成。
- 本章中的航拍与地面候选匹配点通过几何一致性检验与几何模型验证进行过滤。

- 本章中航拍与地面模型通过连接跨越航拍与地面图像的特征点轨迹进行 BA 实现融合。

3.2 方法概述

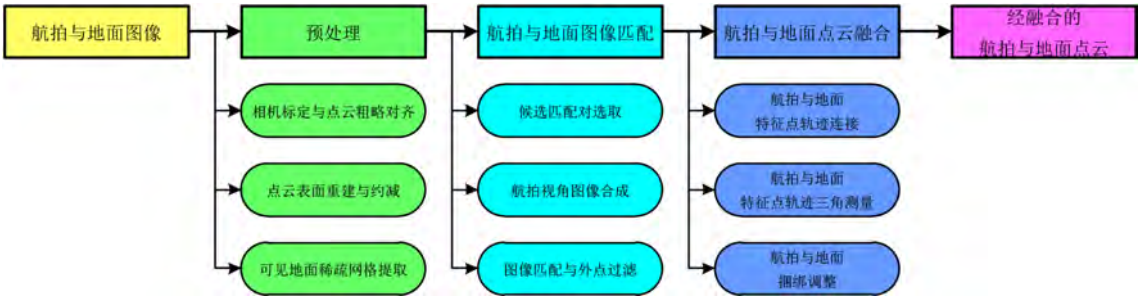

图 3.2: 本章航拍与地面点云融合方法的流程图。该方法主要包含三部分：（1）预处理；（2）航拍与地面图像匹配；（3）航拍与地面点云融合。
Figure 3.2: Pipeline of the reconstruction method by merging ground and aerial point clouds in this chapter. The method contains three main steps: (1) pre-processing; (2) ground-to-aerial image matching; and (3) ground-to-aerial point cloud merging.

本章通过先匹配航拍与地面图像，后融合航拍与地面点云的方式实现建筑场景的完整建模。本章方法的流程图如图3.2所示。该方法的输入和输出分别是航拍与地面图像以及经融合的航拍与地面点云。该方法主要包含三部分：（1）预处理；（2）航拍与地面图像匹配；（3）航拍与地面点云融合。

3.3 预处理

本章方法中预处理包含三个部分：（1）相机及点云的粗略对齐；（2）地面稀疏点云网格化及约减；（3）航拍与地面相机的可见地面稀疏网格提取。预处理步骤的目的是先粗略对齐航拍与地面相机然后提取每个航拍与地面相机的可见地面网格。

3.3.1 相机及点云粗略对齐

本章方法首先对航拍与地面图像分别采用 SfM[73] 获取对应的相机姿态与稀疏点云，然后将航拍与地面的相机及稀疏点云对齐至地理坐标系。在进行点云地理对齐时，本章方法首先将航拍相机及稀疏点云通过相机内置 GPS 置于地理坐标系下，然后人工将地面稀疏点云与航拍稀疏点云粗略对齐。具体来说，本章方法先在已置于地理坐标系下的航拍点云与原始地面点云上选取三对（近似）三维对应点。

然后，采用方法 [35] 计算它们之间的相似变换并将地面点云粗略对齐至航拍点云。需要注意的是，由于经 SfM 生成的航拍与地面点云在噪声程度、密度与精度方面差异明显，因此通过人工对齐的方式很难实现较高精度的对齐。另外，此处粗略对齐的目的是将航拍与地面相机与点云大致置于地理坐标系下，而非精确融合航拍与地面点云。

3.3.2 点云网格化及约减

在相机及点云粗略对齐之后，本章采用方法 [7] 对地面稀疏点云进行表面重建，得到稀疏网格。另外，由于本章的航拍视角图像合成方法并不关注结构细节，且考虑到计算效率的因素，本章采用 QEM 算法 [83] 对地面稀疏网格进行约减。本章中约减比设为 1%。注意，本章方法只对地面稀疏点云进行点云网格化及约减，约减后的地面稀疏网格用做后续图像合成及匹配的媒介。

3.3.3 可见地面网格提取

给出约减后的地面稀疏网格及所有经粗略对齐的航拍与地面相机位姿，本章方法将所有相机的可见网格提取出来。由于本章中的航拍视角图像合成方法基于可见网格中三角面片诱导的单应变换，因此可见网格的提取结果十分重要。给出一个三维网格以及一个相机内外参数，提取该相机的可见网格在计算机图形学领域是一个经典问题，现有方法包括基于深度缓冲器（z-buffer）的方法 [84] 以及基于光线投射（ray-casting）的方法 [85]。在此，本章方法采用的是 OpenMVS 库中的一种基于八叉树的方法实现可见网格的高效提取。

3.4 航拍与地面图像匹配

经过预处理步骤，航拍与地面相机已粗略对齐且每个相机的可见网格均已提取出来。接下来，本章通过如下三个关键子步骤实现航拍与地面图像的匹配：（1）基于公共可见网格的候选匹配对选取；（2）基于网格诱导的航拍视角图像合成；（3）基于几何一致性检验和几何模型验证的匹配外点过滤。

3.4.1 候选匹配对选取

假设共有 N_a 幅航拍图像与 N_g 幅地面图像，如果没有任何先验信息的话，本章方法需要在航拍与地面图像间进行 N_aN_g 次图像匹配。然而，由于在本章中已

对各相机的地面可见网格进行了提取，每对图像的公共可见网格可用作候选图像匹配对选取的依据。接下来将具体介绍本章的基于公共可见网格的候选匹配对选取方法。

对于任意一对航拍与地面图像，本章方法先获取它们可见网格的交集作为初步公共可见网格。然而，通过上述方式获取的公共可见网格中往往存在着一些离散的三角面片，会影响后续的图像合成与匹配。因此，本章方法希望获取初步公共可见网格中的最大连通分量用作图像合成与匹配。具体来说，本章方法首先构建一个 FG，其中图的顶点为各个公共可见面片，边为公共可见面片中的公共边。然后，本章方法提取图中的连通分量。

在每个连通分量中，本章方法将其中的每个面片 f 投影至当前航拍图像上获取其投影面积 S_f 用以对 FG 进行加权。这么做的原因是有着更大 S_f 的面片可以为后续的图像合成与匹配过程提供更多的信息。因此，本章方法将 FG 中的 WLCC 作为最终的公共可见网格。假设共提取得到 N_{cc} 个连通分量，各连通分量分别有 $n_i(i = 1, 2, \ldots, N_{cc})$ 个面片。本章方法通过如下方式将第 i^* 个连通分量选做 WLCC，即最终公共可见网格：

$$i^* = \arg\max \sum_{j=1}^{n_i} S_{f(i,j)} \tag{3.1}$$

其中，$S_{f(i,j)}$ 为第 i 个连通分量中的第 j 个面片在当前航拍图像上的投影面积。在获取公共可见网格后，本章方法将其投影至当前航拍图像并获取其二维包围盒。如果该包围盒足够大（本章中大于 $256px \times 256px$），本章方法认为当前的图像对为候选匹配对。

3.4.2 航拍视角图像合成

理论上来讲，公共可见网格中的每一面片均可以在航拍相机 $C_g(\boldsymbol{K}_a, \boldsymbol{R}_a, \boldsymbol{t}_a)$ 与地面相机 $C_g(\boldsymbol{K}_g, \boldsymbol{R}_g, \boldsymbol{t}_g)$ 之间诱导一个单应，其中，$\boldsymbol{K}, \boldsymbol{R}, \boldsymbol{t}$ 分别为经过粗略对齐的相机内参矩阵，旋转矩阵与平移向量。因此，给出候选匹配对中的一对航拍与地面图像与相应的公共可见网格，航拍视角图像可以通过面片诱导的单应对地面图像进行变换来合成。

对于每个面片，其相应的单应变换可通过 [63] 中介绍的三点法计算得到。首

先，给出的航拍与地面图像之间的基本矩阵计算如下：

$$\boldsymbol{F}_{ag} = \boldsymbol{K}_g^{-T}[\boldsymbol{t}_{ag}]_\times \boldsymbol{R}_{ag}\boldsymbol{K}_a^{-1} \tag{3.2}$$

其中，$\boldsymbol{R}_{ag} = \boldsymbol{R}_g\boldsymbol{R}_a^T$ 与 $\boldsymbol{t}_{ag} = \frac{\boldsymbol{t}_g - \boldsymbol{R}_{ag}\boldsymbol{t}_a}{\|\boldsymbol{t}_g - \boldsymbol{R}_{ag}\boldsymbol{t}_a\|}$ 分别为两相机之间的相对旋转与相对平移。另假设当前面片顶点在航拍与地面图像上的投影分别为 $\boldsymbol{x}_{g(i)}$ 与 $\boldsymbol{x}_{a(i)}, (i = 1, 2, 3)$，那么由该面片诱导的单应可计算如下：

$$\boldsymbol{H}_{ag} = [\boldsymbol{e}_g]_\times \boldsymbol{F}_{ag} - \boldsymbol{e}_g(\boldsymbol{M}^{-1}\boldsymbol{b})^T \tag{3.3}$$

其中，$\boldsymbol{b}$ 为一个三维向量，各分量如下：

$$\boldsymbol{b}_i = \frac{(\boldsymbol{x}_{g(i)} \times ([\boldsymbol{e}_g]_\times \boldsymbol{F}_{ag}\boldsymbol{x}_{a(i)}))^T(\boldsymbol{x}_{g(i)} \times \boldsymbol{e}_g)}{\|\boldsymbol{x}_{g(i)} \times \boldsymbol{e}_g\|^2}, i = 1, 2, 3 \tag{3.4}$$

上式中的 $\boldsymbol{M}$ 为一个行为 $\boldsymbol{x}_{a(i)}^T, (i = 1, 2, 3)$ 的 3×3 矩阵，$\boldsymbol{e}_g$ 为地面图像外极点，由 $\boldsymbol{F}_{ag}^T\boldsymbol{e}_g = \boldsymbol{0}$ 计算得到。

对于每个合成图像中的像素 p 来说，其对应的地面图像像素 p' 的坐标通过三角面片诱导的单应求出。然后，此合成图像中的像素 p 的 RGB 值可通过对地面图像像素 p' 周围像素进行双线性插值给出。图3.3给出了本章航拍视角图像合成的原理示意图。

3.4.3 图像匹配外点过滤

航拍视角图像合成后，航拍图像与合成图像均剪裁至公共可见网格在航拍图像上投影包围盒大小。这是由于只有该包围盒为航拍与地面图像匹配的 RoI。通常，剪裁后的航拍与合成图像通过如下方式匹配：（1）SIFT 特征检测与提取；（2）基于 FLANN 的最近邻搜索；（3）基于 NNDR 验证以及几何模型（如基本矩阵）验证的匹配外点过滤。

然而，尽管地面图像已变换至航拍视角下，航拍图像与合成图像之间仍有较大差异；另外，传统图像匹配方法没有利用一些额外的信息，例如航拍与地面相机已粗略对齐，且合成图像是由地面图像变换至航拍视角下生成的。因此，传统图像匹配方法仅能获取少量匹配点，且匹配外点难以避免。本章借助上述额外信息，提出了一种基于几何一致性检验与几何模型验证的匹配外点过滤方法。需要注意的是，

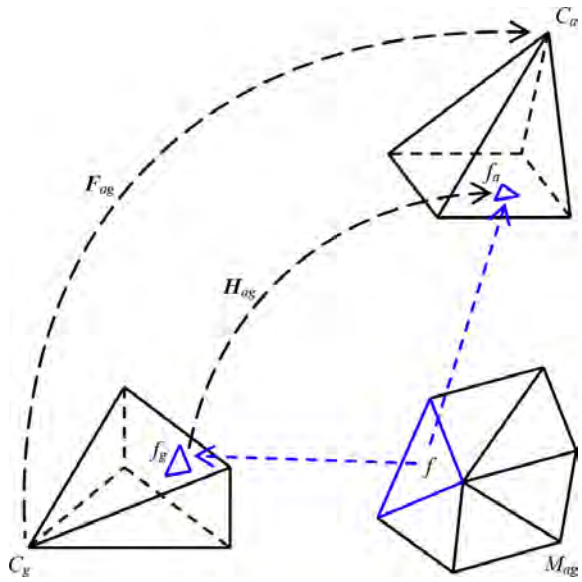

图 3.3: 本章航拍视角图像合成的原理示意图。C_a 与 C_g 为一对航拍与地面图像，$\boldsymbol{F}_{ag}$ 为它们之间的基本矩阵。M_{ag} 为 C_a 与 C_g 的公共可见网格，f 为 M_{ag} 中的一个面片，f_a 与 f_g 分别为 f 在 C_a 与 C_g 上的投影。$\boldsymbol{H}_{ag}$ 为 f 诱导的 f_a 与 f_g 之间的单应变换。需要注意的是，M_{ag} 中的每一个面片均诱导一个单独的单应变换。

Figure 3.3: Schematic diagram of the proposed aerial-view synthesis method in this chapter. C_a and C_g are a pair of aerial and ground cameras, and $\boldsymbol{F}_{ag}$ is the fundamental matrix between them. M_{ag} is the co-visible mesh of C_a and C_g. f is a facet in M_{ag}, and f_a and f_g are the projections of f in C_a and C_g, respectively. $\boldsymbol{H}_{ag}$ is the homography between f_a and f_g induced by the facet f. Note that each facet in M_{ag} induces a unique homography.

在图像检索领域，几何一致性检验 [86, 87] 与几何模型验证 [88, 89] 经常用于过滤候选匹配点。但是由于上述方法中相机位姿未知或部分已知，它们仅在匹配外点过滤时提供弱约束。然而，在本章中由于相机位姿已知且经过粗略对齐，可以采用强约束对航拍与地面匹配外点进行过滤，具体描述如下。

对于一对航拍与地面图像，本章方法提取合成图像的 SIFT 特征并从航拍图像特征数据库中获取航拍图像的 SIFT 特征。由于航拍图像无需变换，用于此处的航拍图像特征数据库为 SfM 阶段（3.3.1节）中生成的。然后，本章方法通过 FLANN 获取航拍与合成图像间的候选匹配点，记为 $\{\boldsymbol{x}^{pt}_{a(i)}, \boldsymbol{x}^{pt}_{g(i)}\}, (i = 1, 2, \ldots, n_{pt})$。通过本章的图像合成方法，通常一些合成图像中特征点的几何特性，例如尺度 σ 与主方向 θ，与航拍图像中的匹配特征点是接近的。因此，本章方法采用类似2.4.3节中式2.11的方式对候选匹配点进行过滤，只保留满足几何一致性约束的匹配点。通过上述过滤的匹配点记为 $\{\boldsymbol{x}^{fl}_{a(i)}, \boldsymbol{x}^{fl}_{g(i)}\}, (i = 1, 2, \ldots, n_{fl})$。

另外，如果物体的深度变化相对于其到相机的距离很小，物体在该相机上的投

影可以近似看作仿射投影 [63]。由于航拍相机通常距地面较远，航拍相机中的投影满足上述假设。假设在此将给出的航拍相机的投影矩阵记为：

$$\boldsymbol{P}=\begin{pmatrix}\boldsymbol{P}_{2\times3} & \boldsymbol{p}_{2\times1}\\ \boldsymbol{p}_{1\times3} & 1\end{pmatrix} \tag{3.5}$$

其中，由于航拍相机中的投影近似仿射投影，$\boldsymbol{p}_{1\times3}\to\mathbf{0}^T$。另外，由于航拍与地面点云已经过粗略对齐，它们之间的相似变换趋近于不变，即 $s\to1$，$\boldsymbol{R}\to\boldsymbol{I}$ 以及 $\boldsymbol{t}\to\mathbf{0}$。因此，地面点云相对于航拍相机的投影也近似为仿射投影：

$$\boldsymbol{P}'=\boldsymbol{P}\boldsymbol{T}=\begin{pmatrix}\boldsymbol{P}_{2\times3} & \boldsymbol{p}_{2\times1}\\ \boldsymbol{p}_{1\times3} & 1\end{pmatrix}\begin{pmatrix}s\boldsymbol{R} & \boldsymbol{t}\\ \mathbf{0}^T & 1\end{pmatrix}=\begin{pmatrix}s\boldsymbol{P}_{2\times3}\boldsymbol{R} & \boldsymbol{P}_{2\times3}\boldsymbol{t}+\boldsymbol{p}_{2\times1}\\ s\boldsymbol{p}_{1\times3}\boldsymbol{R} & \boldsymbol{p}_{1\times3}\boldsymbol{t}+1\end{pmatrix} \tag{3.6}$$

这是因为 $\frac{s\boldsymbol{p}_{1\times3}\boldsymbol{R}}{\boldsymbol{p}_{1\times3}\boldsymbol{t}+1}\to\mathbf{0}^T$（由于 $s\boldsymbol{p}_{1\times3}\boldsymbol{R}\to\mathbf{0}^T$ 以及 $\boldsymbol{p}_{1\times3}\boldsymbol{t}+1\to1$）。理论上来讲，同一个三维空间点集的两个二维仿射投影点集之间的变换为一个二维仿射变换。在此需要注意的是，通过同一个投影矩阵，定义于式3.5中的 $\boldsymbol{P}$，对航拍与地面点云，$\boldsymbol{X}_a$ 与 $\boldsymbol{X}_g$，进行投影本质上等价于通过两个不同的投影矩阵，定义于式3.5中的 $\boldsymbol{P}$ 与定义于式3.6中的 $\boldsymbol{P}'$，对航拍点云 $\boldsymbol{X}_a$ 分别进行投影。这是由于在理想状态下 $\boldsymbol{X}_a=\boldsymbol{T}\boldsymbol{X}_g$。

基于上述分析，可对获取的匹配点进行进一步过滤。给出经过几何一致性检验的匹配点 $\{\boldsymbol{x}^{fl}_{a(i)},\boldsymbol{x}^{fl}_{g(i)}\},(i=1,2,\ldots,n_{fl})$，本章方法通过利用 RANSAC 估计它们之间的仿射变换的方式实现匹配点的过滤。具体来说，给出需要进一步过滤的匹配点，本章方法首选随机选取用于估计二维仿射变换的最小集合（大小为 3），然后采用 DLT 算法 [63] 进行仿射变换估计并设置距离阈值为 $4px$ 以获取对应的内点集合。上述过程重复多次（本章中重复 100 次）以获取有着最多内点数量的最大一致集。这些内点即为过滤后的匹配点，将其记为 $\{\boldsymbol{x}^{af}_{g(i)},\boldsymbol{x}^{af}_{a(i)}\},(i=1,2,\ldots,n_{af})$。在此需要强调的是，本章的基于仿射变换的方法借助点到点约束实现匹配点的过滤，不像基于基本矩阵或者本质矩阵的方法，借助的是点到线约束实现匹配点的过滤。因此，本章方法对于匹配外点滤除更为有效。

然后，本章方法将匹配点坐标转换到原始图像坐标系下，记做 $\{\tilde{\boldsymbol{x}}^{af}_{g(i)},\tilde{\boldsymbol{x}}^{af}_{a(i)}\},(i=1,2,\ldots,n_{af})$。该匹配点即为最终的航拍与地面图像匹配结果。如果 $n_{af}\geq n_{th}$，其中 n_{th} 为匹配点数量阈值且在本章中设为 16，本章中认为当前的航拍与地面图像对是匹配的。图3.4给出了一对航拍与地面图像匹配结果示例。如图3.4所示，原始

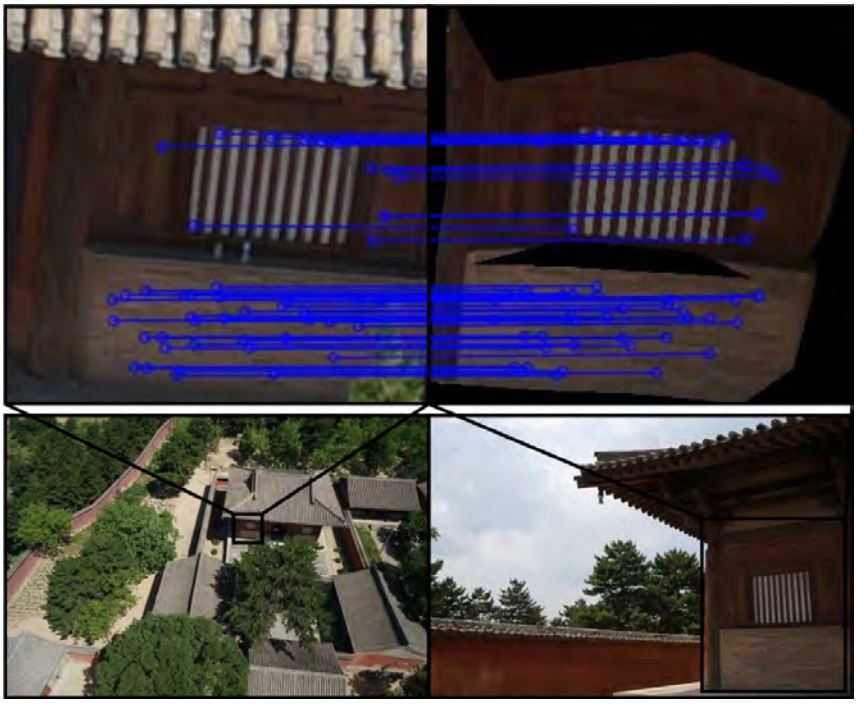

图 3.4: 一对航拍与地面图像匹配结果示例。第一行为航拍与合成图像公共区域的匹配结果，其中蓝色线段表示匹配点。第二行为原始航拍与地面图像对，其中黑色矩形表示航拍与地面图像公共可见区域。
Figure 3.4: An example of ground-to-aerial image matching result. The first row is the matching result between the co-visible regions of the aerial and synthetic images, where the blue segments denote the point matches. The second row is the original aerial and ground image matching pair, where the black rectangles denote the co-visible regions for image matching.

航拍与地面图像之间在视角与尺度方面差异巨大，然而这些差异在将地面图像进行航拍视角合成后很大程度上消除了。因此，本章方法在图像对公共区域上获取了许多匹配点。

3.5 航拍与地面点云融合

在获取航拍与地面图像之间的二维匹配点后，一个最直接的方式是通过估计航拍与地面点云之间的相似变换实现点云的对齐。该相似变换通过将二维匹配点反投影至点云获取的三维对应点计算得到。上述过程可通过将视线与空间连续的航拍与地面稀疏网格分别相交获取。该视线由航拍或地面相机光心射出，指向对应的匹配特征点。然而，上述点云对齐方法有两个主要的缺点：（1）通常情况下由图像重建得到的点云与网格噪声较大，因此得到的三维对应点的精度较低；（2）基于图像的重建结果中累积误差与场景飘移是普遍存在的，仅通过一个相似变换不

足以表征航拍与地面点云之间的变换关系。因此，本章通过 BA 实现航拍与地面点云的融合。

正如3.3.1节所述，航拍与地面点云均重建于它们的局部坐标系下。为此，本章方法需要通过连接航拍与地面特征点轨迹的方式将跨越航拍与地面点云的约束引入 BA 中实现点云的融合。根据3.4.3节中的介绍，在进行航拍与地面图像匹配时，航拍图像特征点是由航拍图像特征数据库中获取的。因此，本章方法可以得知航拍与地面图像匹配点中的航拍特征点属于哪些原始航拍特征点轨迹。然后，本章方法就可以将航拍与地面匹配点连入原始航拍特征点轨迹。

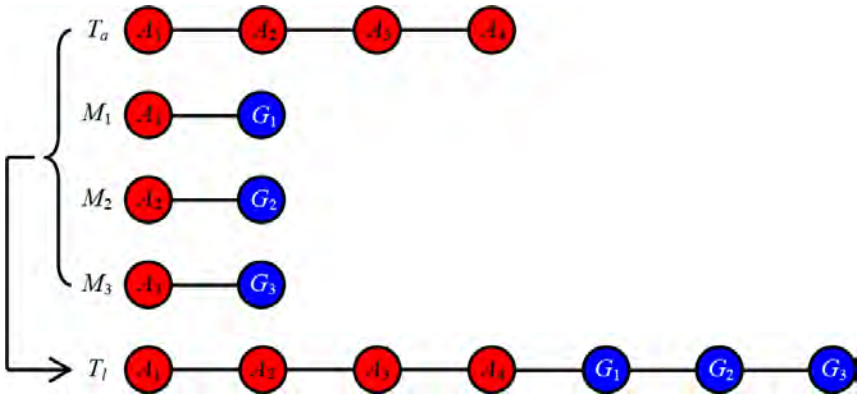

图 3.5: 航拍与地面特征点轨迹连接示意图。$A_i,(i=1,2,3,4)$ 为四个航拍特征点，$G_i,(i=1,2,3)$ 为三个地面特征点。T_a 为一个原始航拍特征点轨迹，$M_i,(i=1,2,3)$ 为三对航拍与地面匹配点。T_l 为一个将 $M_i,(i=1,2,3)$ 连入 T_a 的新特征点轨迹。
Figure 3.5: Schematic diagram of ground-to-aerial track linking. $A_i,(i=1,2,3,4)$ are 4 aerial feature points and $G_i,(i=1,2,3)$ are 3 ground feature points. T_a is the original aerial track, $M_i,(i=1,2,3)$ are 3 pairs of ground-to-aerial point matches. T_l is a new track by linking $M_i,(i=1,2,3)$ to T_a.

图3.5为一个航拍与地面特征点轨迹连接示意图。T_a 为一个原始航拍特征点轨迹，$M_i,(i=1,2,3)$ 为三对航拍与地面匹配点。由于可以从 T_a 中找到上述三对匹配点中的航拍特征点，本章方法将匹配点中的地面特征点连入 T_a 中生成新的特征点轨迹 T_l，用于后续的 BA。图3.6为一个连接后的航拍与地面特征点轨迹示例。需要注意的是，由于地面图像经过变换且在进行图像匹配时在合成图像上重新提取了特征点，因此在进行航拍与地面特征点轨迹连接时，并未涉及到原始地面特征点轨迹。

在将航拍与地面匹配点连入原始航拍特征点轨迹后，本章方法对连接的航拍与地面特征点轨迹进行三角测量以获取其空间坐标，用做后续 BA 初值。在此采用了一种类似于方法 [9] 的基于 RANSAC 的三角测量法。但是，在每次随机选取

图 3.6: 航拍与地面特征点轨迹连接示例。第一行为三个航拍图像块与三个地面图像块，其中蓝色线段表示跨越图像的特征点轨迹。第二行为原始的航拍与地面图像，其中黑色矩形标示出了第一行中的图像块。

Figure 3.6: An example of linked ground-to-aerial track. The first row contains three aerial and three ground image patches, where the blue segment denotes the linked track across views. The second row contains original aerial and ground images, where the black rectangles denote the image patches in the first row.

最小集合时，本章方法限制选取的两个投影点分属于航拍与地面图像以保证能够引入航拍与地面之间的约束关系。

随后，本章方法通过如下的全局 BA 实现航拍与地面点云的融合：

$$\min_{\boldsymbol{R}_i,\boldsymbol{t}_i,\boldsymbol{X}_j} \sum_{i,j} \delta_{ij} \|\boldsymbol{x}_{ij} - \gamma(\boldsymbol{K}_g, \boldsymbol{K}_a, \boldsymbol{R}_i, \boldsymbol{t}_i, \boldsymbol{X}_j)\|_{huber}, \tag{3.7}$$

其中，$\boldsymbol{x}_{ij}$ 为特征点轨迹中的观测的投影点；$\boldsymbol{K}_a$ 与 $\boldsymbol{K}_g$ 为航拍与地面相机内参矩阵；$\boldsymbol{R}_i$ 与 $\boldsymbol{t}_i$ 为（航拍与地面）相机的旋转矩阵与平移向量；$\boldsymbol{X}_j$ 为所有待优化的空间点，包括航拍与地面稀疏点云中的所有点以及由连接的航拍与地面特征点轨迹三角测量得到的点；$\gamma(\cdot)$ 为投影方程；δ_{ij} 为符号函数：若相机 i 可以观测到点 j，$\delta_{ij} = 1$，否则，$\delta_{ij} = 0$。另外，针对不可避免的航拍与地面图像误匹配问题，上式引入和 Huber 损失函数。需要注意的是，由于所有（航拍/地面）图像均采用同一（航拍/地面）相机拍摄，因此它们共享同一相机内参矩阵。另外，本章方法认为经过3.3.1节中的 SfM 过程，航拍与地面相机内参数均已精确标定，因此它们在此处进行 BA 是保持不变。本章采用一个用于建模、求解大型、复杂优化问题的开源库 Ceres Solver 对上式中定义的问题进行求解。由于 BA 的过程可能会改变点云与相机所在的坐标系，因此在 BA 之后本章方法重新将融合后的航拍与地面点云以及对应的相机置于地理坐标系下。

3.6 实验结果

接下来，本节对本章中的融合航拍与地面点云的场景重建方法进行了测评。本节首先介绍用于进行方法评测的四个数据集，然后，在评测数据集上，本节对本章提出的航拍与地面的图像匹配及点云融合方法进行了评测。

3.6.1 数据集

表 3.1: 用于方法测评的数据集元数据。
Table 3.1: Meta-data of the datasets for method evaluation.

数据集	EMH	MJH	NCT	FGT
覆盖面积	$600m^2$	$500m^2$	$3100m^2$	$34000m^2$
航拍图像数量：N_a	208	579	772	1596
地面图像数量：N_g	2439	2619	2790	6978
航拍图像空间分辨率	$14.29mm/px$	$16.07mm/px$	$8.47mm/px$	$15.96mm/px$
地面图像空间分辨率	$0.77mm/px$	$0.97mm/px$	$0.73mm/px$	$0.88mm/px$
航拍稀疏点云点数	$0.69M$	$1.56M$	$1.15M$	$4.81M$
航拍稀疏点云点数	$0.79M$	$0.71M$	$1.54M$	$2.36M$
地面稀疏网格面片数	$0.46M$	$0.35M$	$0.94M$	$1.31M$

与第2章类似，本节仍在自建数据集上进行方法评测。本章的测试数据集包括两个中国古代佛殿东大殿（EMH）与文殊殿（MJH）以及两个中国古代寺庙南禅寺（NCT）与佛光寺（FGT）。上述建筑均为典型的中国古代建筑且在绝大多数情况下，一个寺庙包含有着特定布局的若干佛殿。因此，寺庙的覆盖面积大于佛殿。各数据集的元数据在表3.1中给出。表3.1中最后一行为约减前的地面稀疏网格数量，因此给出约减比（本章中为 1%），即可获知约减后的网格数量。另外，由于本章方法没有对航拍稀疏点云进行网格化处理，因此表3.1中没有航拍稀疏网格这一项。如表3.1所示，航拍与地面图像在尺度（空间分辨率）上差异明显。另外，图3.8的第一列给出了一些航拍与地面图像示例，通过这些示例可知两类图像在视角与尺度方面存在很大差异。

3.6.2 航拍与地面图像匹配结果

首先，本节在四个数据集上进行了候选匹配对选取实验，结果如表3.2所示。由表3.2可知本章中的候选匹配对选取方法大幅度减少了待匹配的图像对数，进而有效提升了后续航拍与地面图像匹配的效率。

表 3.2: 候选匹配对选取及图像匹配结果。其中，候选匹配对为所有可能进行匹配的图像对；选取匹配对为通过匹配对选取方法获取的图像对；匹配图像对为可通过航拍与地面图像匹配方法实现匹配的图像对。
Table 3.2: Candidate matching pairs selection and image matching results. Candidate matching pairs are all possible image pairs for matching. Selected matching pairs are the image pairs selected by the proposed candidate matching pairs selection method. Matched image pairs are the image pairs matched by the proposed ground-to-aerial image matching method.

数据集	EMH	MJH	NCT	FGT
候选匹配对数量：N_gN_a	$0.51M$	$1.52M$	$2.15M$	$11.14M$
选取匹配对数量：N_s	$7.26K$	$14.90K$	$6.50K$	$45.24K$
约减比 $\frac{N_s}{N_gN_a}$	1.43%	0.98%	0.34%	0.41%
匹配图像对数量：N_m	$2.04K$	$5.17K$	$2.46K$	$7.23K$

表 3.3: 本章的航拍与地面图像匹配方法与其它对比方法在匹配对类型，特征类型以及外点过滤方式上的区别。
Table 3.3: Differences between the proposed ground-to-aerial image matching method in this chapter and other four methods for comparison in terms of matching pair type, feature type and outlier filtering scheme.

匹配方法	匹配对类型	特征类型	外点过滤方式
本章方法	航拍与合成图像	SIFT	几何一致性检验及仿射变换验证
RepMatch	航拍与地面图像	ASIFT	匹配一致性检验及外极几何引导过滤
ASIFT	航拍与地面图像	ASIFT	NNDR 及基本矩阵验证
Warp+SIFT	航拍与合成图像	SIFT	NNDR 及基本矩阵验证
SIFT	航拍与地面图像	SIFT	NNDR 及基本矩阵验证

基于选取的候选匹配对，本节在测评数据集上对本章的航拍与地面图像匹配方法的召回率、精度以及效率进行了测评，并与其它四种方法进行了对比。对比方法包括 SIFT、Warp+SIFT、ASIFT 以及 RepMatch[90]。表3.3给出了本章方法与其它四种对比方法之间的区别。其中，合成图像是根据本章中的航拍视角图像合成方法由地面图像生成。另外，由于直接匹配原始航拍与地面图像结果总是失败。为简化该问题，本节首先根据公共可见网格获取航拍与地面图像的公共可见区域并对图像进行剪裁，然后将剪裁后的地面图像降采样至剪裁后的航拍图像分辨率。上述操作的目的是为了消除两种图像之间的尺度差异并且去除图像中的无关信息。

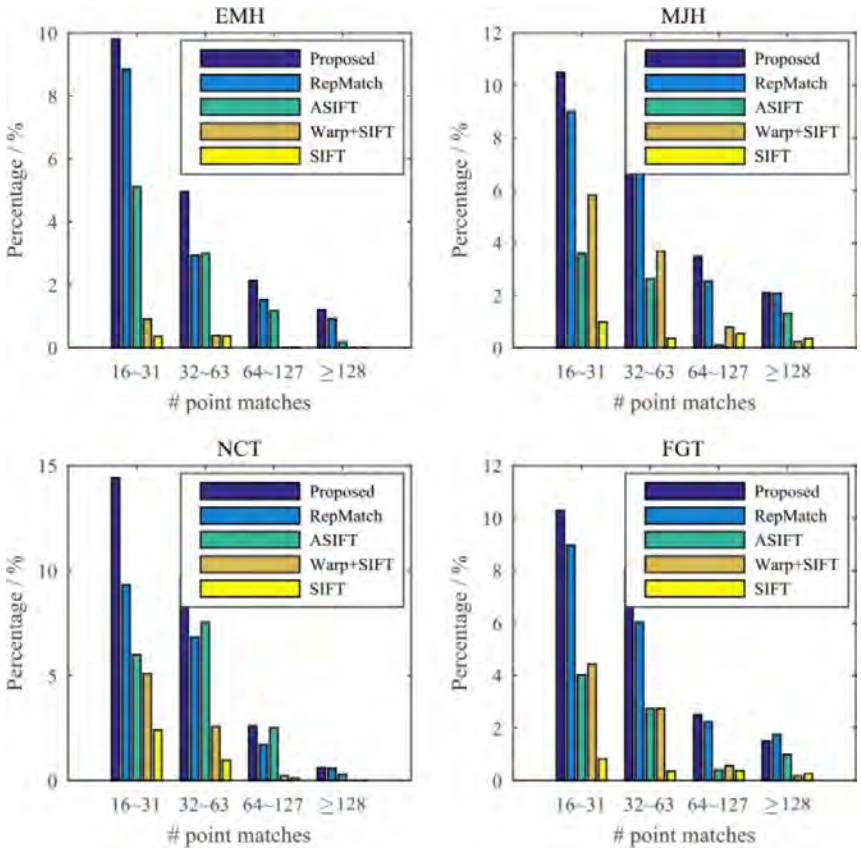

图 3.7: 航拍与地面图像匹配对比结果：召回率。图中 y 轴表示匹配图像对数占选取图像对数的百分比，x 轴表示匹配内点数量区间。

Figure 3.7: Ground-to-aerial image matching results: recall rate. y-axis is the ratio in percentage form of the number of matched image pairs to the number of selected image matching candidate pairs, and x-axis is the interval of point matches number.

3.6.2.1 召回率

由于没有测评航拍与地面图像匹配结果的真值，本节将通过不同方法得到的匹配图像对数 N_m（表3.2最后一行）以及每对图像匹配点数 $n_{af(i)}(i = 1, 2, \ldots, N_m)$ 用做衡量匹配方法召回率的指标。这是因为图像匹配对数以及匹配特征点数越多，能够召回更多正确匹配点的可能性就越大。实验结果以直方图的形式示于图3.7。图3.7中 y 轴表示匹配图像对数占选取图像对数 N_s（表3.2第二行）的百分比，x 轴表示匹配内点数量区间。如图3.7所示，本章方法的匹配图像对数及图像匹配点数均多于其它四种对比方法。在其它四种对比方法中，RepMatch[90] 效果最好。本章方法与 RepMatch[90] 在召回率方面性能较好的可能原因是这两种方法均没有 NNDR 验证步骤，而 NNDR 验证存在易舍弃正确匹配点的问题。

3.6.2.2 精度

表 3.4: 航拍与地面图像匹配对比结果：精度与效率。
Table 3.4: Ground-to-aerial image matching results: precision and efficiency.

数据集	本章方法		RepMatch		ASIFT		Warp+SIFT		SIFT	
	精度	效率	精度	效率	精度	效率	精度	效率	精度	效率
EMH	98.29%	1.12s	93.85%	8.53s	89.79%	38.37s	85.29%	0.96s	78.89%	0.10s
MJH	98.59%	1.03s	92.68%	8.03s	87.05%	39.34s	88.11%	0.91s	74.19%	0.12s
NCT	98.14%	0.90s	92.83%	8.34s	87.51%	36.31s	85.06%	0.79s	75.71%	0.08s
FGT	98.50%	0.96s	93.11%	8.27s	87.83%	38.02s	86.31%	0.93s	76.52%	0.11s

接下来，本节通过人工验证得到的匹配点是否为正确的匹配点的方式分析不同方法的匹配精度。然而，由于匹配的图像对数过多，例如本章匹配方法在 FGT 数据集上得到了 7225 对匹配对。因此从所有图像匹配对中辨别所有的正确匹配点十分耗时且不切实际。因此，本节只对其中的一个随机抽取的子集进行人工验证。具体来说，本节首先从各个数据集中抽取 10 对各个对比方法均可匹配的航拍与地面图像对，然后人工获取抽取图像对的匹配精度并将结果列于表3.4中。如表3.4所示，对于匹配精度来说，本章方法 $>$ RepMatch[90] $>$ ASIFT $\approx$ Warp+SIFT $>$ SIFT。上述结果说明本章中的基于几何一致性检验与几何模型验证的匹配外点过滤方法十分有效。由于 RepMatch[90] 引入了匹配一致性检验与外极几何引导过滤，该方法同样取得了较好的匹配精度。这里的匹配一致性指的是如下三个特性的联合度量：匹配密度，平滑性与空间覆盖范围。另外，由于 ASIFT 与 Warp+SIFT 中均包含应对视角变化的图像变换操作，这两种方法匹配精度类似且高于 SIFT。

3.6.2.3 效率

最后，本节通过匹配每对图像的平均耗时衡量各匹配算法的效率，实验结果仍列于表3.4。如表3.4所示，在匹配效率方面，SIFT $>$ Warp+SIFT $\approx$ 本章方法 $>$ RepMatch[90] $>$ ASIFT。由于存在较复杂的操作，RepMatch[90] 与 ASIFT 效率较低：RepMatch[90] 基于匹配一致性训练了一个分类函数；对于 ASIFT 来说，为找到相似视角，该方法进行了多次图像变换。在所有方法中，SIFT 匹配效率最高，然而由于 SIFT 不能应对航拍与地面图像间的视角差异，匹配结果（召回率、精度）较差。另外，由于本章方法中包含一个基于 RANSAC 的仿射变换估计步骤，其效率略低于 Warp+SIFT。

综上所述，本章的航拍与地面图像匹配方法在召回率与精度方面表现最好且有着较高的效率。

3.6.3 航拍与地面点云融合结果

接下来，本节在四组测评数据集上对本章的航拍与地面点云融合方法进行了测试。首先，本节通过点云融合定性结果验证方法的有效性。其次，3.3节中的两个操作，粗略对齐以及网格约减，可能会影响点云融合的结果。因此，本节对本章点云融合方法对粗略对齐以及网格约减比的依赖性进行了评测。最后，本节将本章点云融合方法与其它方法进行了定量对比。

为定量评测本章点云融合方法精度，本节采用了文献 [34] 中的一个近似评测方法。对于地面点云中的每个点 $\boldsymbol{X}_g$，其法向记为 $\boldsymbol{n}_g$，本节首先在航拍点云中找到它的最近点 $\boldsymbol{X}_a$。然后，本节计算 $\boldsymbol{X}_g$ 到 $\boldsymbol{X}_a$ 的沿着 $\boldsymbol{n}_g$ 的投影距离 $d = \frac{\boldsymbol{n}_g \cdot (\boldsymbol{X}_g - \boldsymbol{X}_a)}{\|\boldsymbol{X}_g - \boldsymbol{X}_a\|}$。最后，本节统计地面点云中 d 小于 $n\sigma$ 的点的百分比并绘成 CED 曲线 [11]。本节认为投影距离大于 10σ 的点为同一类点，将它们累加在 11σ 处。尽管上述评测方法不是准确的度量，它仍可以对点云融合精度提供一个相对有效的评测 [34]。注意，在进行精度评测时，并不涉及到连接的航拍与地面特征点轨迹。因此，可以采用上述评测方法定量对比本章方法与基于相似变换估计的点云对齐方法的精度。下文（3.6.3.2节至3.6.3.4节）中的结果均通过该评测方法得到。

3.6.3.1 点云融合定性结果

本章的航拍与地面点云融合算法在四组评测数据上的定性结果如图3.8所示。另外，本节对融合后的点云进行了表面重建，重建得到的网格示于图3.8的最后一列。由融合点云以及重建网格可知，点云较好的融合在了一起，并且通过点云融合，本节可获取完整且细节丰富的建筑场景重建结果。

3.6.3.2 对粗略对齐的依赖性

本节在 EMH 数据集上测试了本章点云融合方法对粗略对齐精度的依赖性。如3.3.1所述，粗略对齐仅需航拍与地面点云中的三对对应点。在此，本节人工选取十对对应点，然后每次从这十对对应点中随机选取三对用于将地面点云粗略对齐至处于地理坐标系下的航拍点云，共进行五次。由于选取的十对对应点精度不同，五次粗略对齐的精度也不同，正如图3.9a 所示。然后，本节采用本章的点云融合算

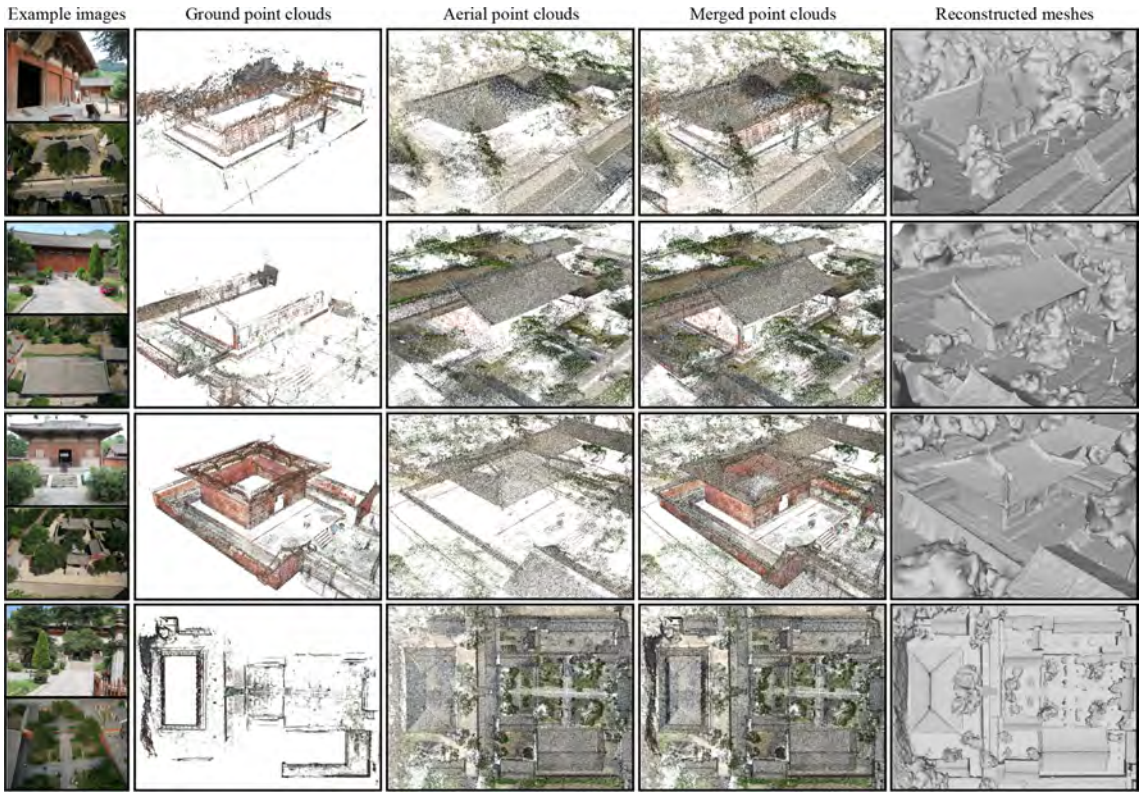

图 3.8: 本章的航拍与地面点云融合算法定性结果。从上到下：在 EMH、MJH、NCT 以及 FGT 数据集上的结果。从左到右：示例航拍与地面图像，地面点云，航拍点云，融合点云以及对融合点云表面重建得到的网格。
Figure 3.8: The qualitative ground-to-aerial point cloud merging results of the proposed method. From top to bottom: the results on EMH, MJH, NCT and FGT datasets. From left to right: example ground and aerial images, ground point clouds, aerial point clouds, merged point clouds, and meshes reconstructed from the merged point clouds.

法在上述五次粗略对齐结果的基础上进行点云融合，融合结果如图3.9b 所示。由图3.9b 可知，尽管粗略对齐精度不同，点云融合精度却能基本保持不变，这说明本章的点云融合方法对粗略对齐结果并不敏感。

3.6.3.3 对网格约减比的依赖性

另外，本节还在 EMH 数据集上测试了本章点云融合方法对网格约减比的依赖性。首先，本节给定不同约减比（0.01%，0.1%，0.5%，1% 以及 2%）对地面稀疏网格进行约减。然后，基于约减后的网格，本节进行了航拍与地面图像匹配以及点云融合。在不同网格约减比下的点云融合结果如图3.10a 所示。由图3.10a 可知，当简约比很低的时候点云融合结果较差。当约减比为 0.01% 时，点云融合结果相对于粗略对齐结果几乎没有变化，这意味着此时点云融合方法没起作用。这是因为在如此之低的约减比下，基于边收缩的 QEM 算法 [83] 不能够输入地面稀疏网格的

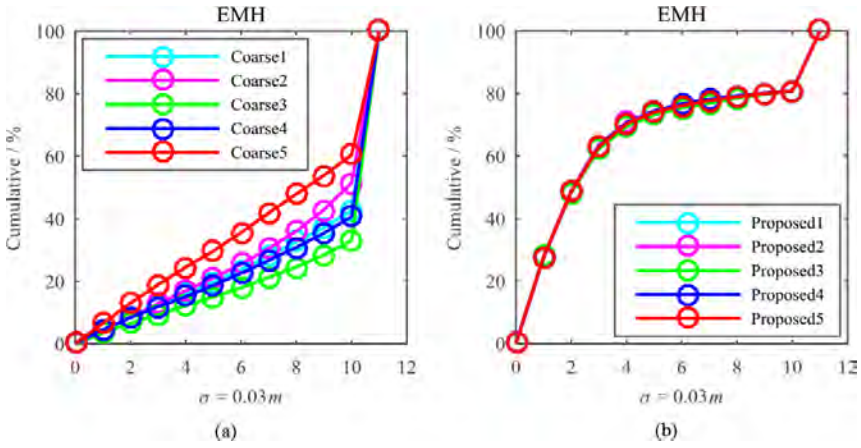

图 3.9: EMH 数据集上的本章点云融合方法对粗略对齐精度的依赖性的实验结果。图（a）中的 Coarse$i, (i = 1, 2, \ldots, 5)$ 与图（b）中的 Proposed$i, (i = 1, 2, \ldots, 5)$ 分别表示五次粗略对齐与点云融合结果。图中的曲线为累积误差分布曲线，其中 y 轴为地面点中最近投影距离小于 $n\sigma$ 占所有地面点的百分比。

Figure 3.9: The result of dependency on coarse alignment accuracy on EMH dataset. Coarse$i, (i = 1, 2, \ldots, 5)$ in (a) and Proposed$i, (i = 1, 2, \ldots, 5)$ in (b) are the 5 trials of coarse alignment results and proposed ground-to-aerial point cloud merging results, respectively. The curves are the cumulative error distributions with the y-axis being the percentage of the ground points with errors $< n\sigma$.

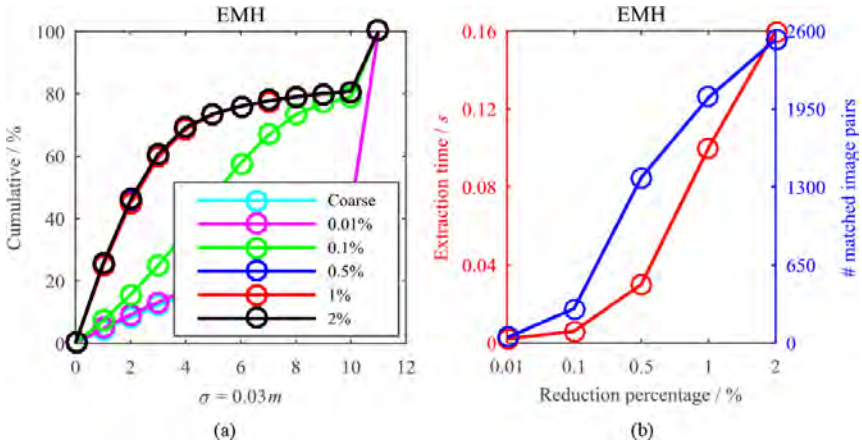

图 3.10: EMH 数据集上的本章点云融合方法对网格约减比的依赖性的实验结果。图（a）为粗略对齐以及在不同网格约减比下的点云融合结果。图（a）中的曲线为累积误差分布曲线，其中 y 轴为地面点中最近投影距离小于 $n\sigma$ 占所有地面点的百分比。图（b）为在不同网格约减比下提取可见网格平均时间以及匹配的图像对数。

Figure 3.10: The result of dependency on mesh reduction percentage on EMH dataset. (a) The coarse alignment result and proposed point cloud merging results with different mesh reduction percentages. The curves are the cumulative error distributions with the y-axis being the percentage of the ground points with errors $< n\sigma$. (b) The average time for visible mesh extraction and the number of matched image pairs with different mesh reduction percentages.

拓扑结构。随着约减比的增大，点云融合结果逐渐变好最后趋于稳定。由上述结果可知，当网格约减比在一个合适的范围内时，本章的点云融合方法对其并不敏感。另外，如图3.10b 所示，随着网格约减比的增大，匹配图像对数与可见网格提取耗时均大幅度增加。因此，为平衡方法的精度与效率，本节将约减比的值设为 1%。

3.6.3.4 点云融合定量对比结果

本节对本章的点云融合方法与与其他三个点云对齐方法进行了比较：Back-proj，Shan et al.[4] 以及 Zhou et al.[34]。上述三种方法通过估计航拍与地面点云之间的相似变换实现点云对齐。对于 Back-proj，用于实现点云对齐的相似变换是由航拍与地面点云的三维对应点通过 RANSAC 估计得到的。该三维对应点是根据航拍与地面二维匹配点对应的视线分别于航拍与地面稀疏网格相交获取，而二维匹配点是根据本章的航拍与地面图像匹配方法获取的。另外，方法 Shan et al.[4] 中用到的稠密点云是通过方法 [75] 得到的；方法 Zhou et al.[34] 中用到的网格是通过方法 [7] 得到的。需要注意的是，上述所有对比方法均需要借助噪声较大的三维几何体（点云或者网格）获取三维对应点，而本章中的点云融合方法仅需要二维航拍与地面匹配点即可。

3.3.1节中的粗略对齐结果以及上述四个方法（本章方法与三个对比方法）结果如图3.11所示。由图3.11可知粗略对齐结果的精度很差，另外四种方法均可在此基础上提升较多的精度。另外，本章的点云融合方法在所有对比方法中精度最高。可能的原因如下：（1）本章方法利用二维图像匹配点，而不是噪声较大的三维空间对应点，实现点云的融合；（2）其它三个对比方法中用到的单个相似变换不足以表征航拍与地面点云之间的变换关系，这种情况在大规模建筑场景中表现尤为明显。在效率方面，相比于稠密重建（采用方法 [75] 在 FGT 数据集上大概需要 30 小时），在稀疏点云网格化上消耗的时间（采用方法 [7] 在 FGT 数据集上大概需要 6 分钟）几乎可以忽略不计。因此，本章中的点云融合方法比基于稠密点云的方法效率更高。

3.6.4 对场景漂移的影响

理论上讲，航拍与地面点云及相机之间的变换关系可通过单个相似变换进行表征。然而，由于基于图像的重建方法存在累积误差与场景漂移的问题，且在大规模场景下尤为明显，上述假设通常不成立。本章的解决方法是通过 BA 的方式实

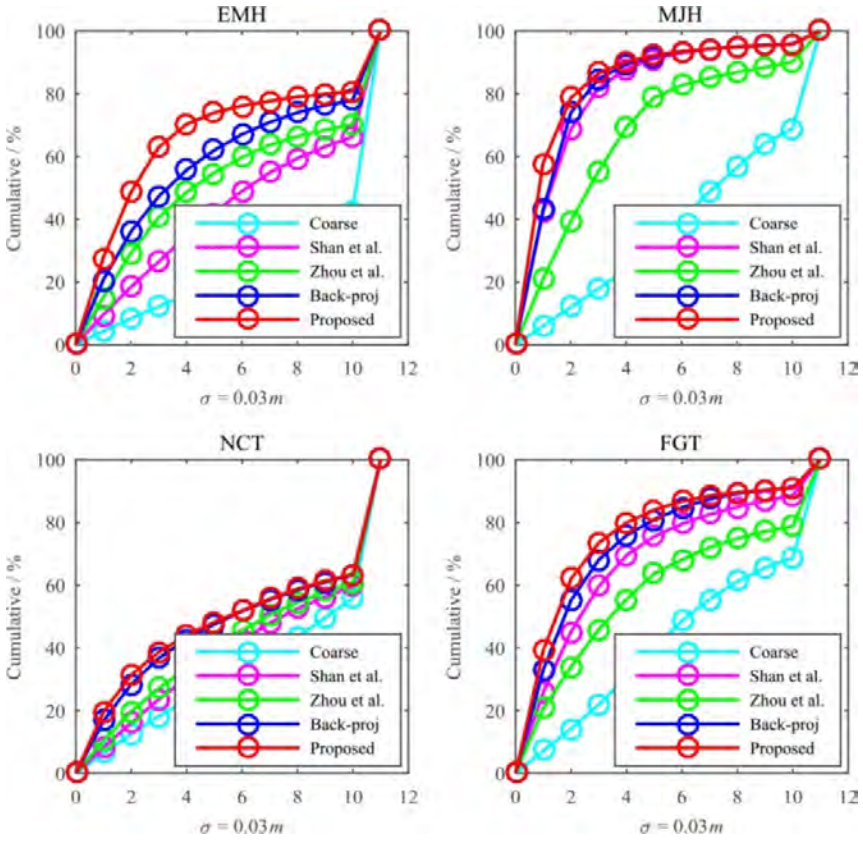

图 3.11: 定量点云融合结果。图中的曲线为累积误差分布曲线，其中 y 轴为地面点中最近投影距离小于 $n\sigma$ 占所有地面点的百分比。

Figure 3.11: The quantitative ground-to-aerial point cloud merging results. The curves are the cumulative error distributions with the y-axis being the percentage of the ground points with errors $< n\sigma$.

现点云的融合，而经融合后的点云与原始的航拍与地面点云（及相机）之间的关系不再是单个相似变换了。为展示原始航拍与地面点云中的场景漂移情况，本节对经 BA 融合前后的相机位姿（位置与朝向）进行了比较。以航拍相机为例，假设在 BA 前后相机旋转平移分别记为 $\boldsymbol{R}_i, \boldsymbol{c}_i$ 与 $\boldsymbol{R}_i^{BA}, \boldsymbol{c}_i^{BA}$。这里需要注意的是 BA 可能会改变相机所处的坐标系，因此需要将 BA 后的相机位姿变换至原始相机所处的坐标系之下。上述变换记做 $\boldsymbol{T}' = \begin{pmatrix} s'\boldsymbol{R}' & \boldsymbol{t}' \\ \boldsymbol{0}^T & 1 \end{pmatrix}$，由 RANSAC 估计得到。然后，相机旋转与位置差异分别通过如下方式计算：$\delta_{\boldsymbol{R}_i} = acos(\|\boldsymbol{R}_i - \boldsymbol{R}_i^{BA}\boldsymbol{R}'^T\|_F)$ 以及 $\delta_{\boldsymbol{c}_i} = \|\boldsymbol{c}_i - \boldsymbol{T}'\boldsymbol{c}_i^{BA}\|$。航拍相机的位姿差异也采用同样的方式进行计算。

本节在 NCT 与 FGT 数据集上进行了上述实验，并将相机旋转与位置差异的累积误差分布曲线绘于图3.12中。若重建结果不存在场景飘移现象，位姿差异应该接近于零。然而，如图3.12所示，相机位姿差异的情况确实存在，且在更大规模的

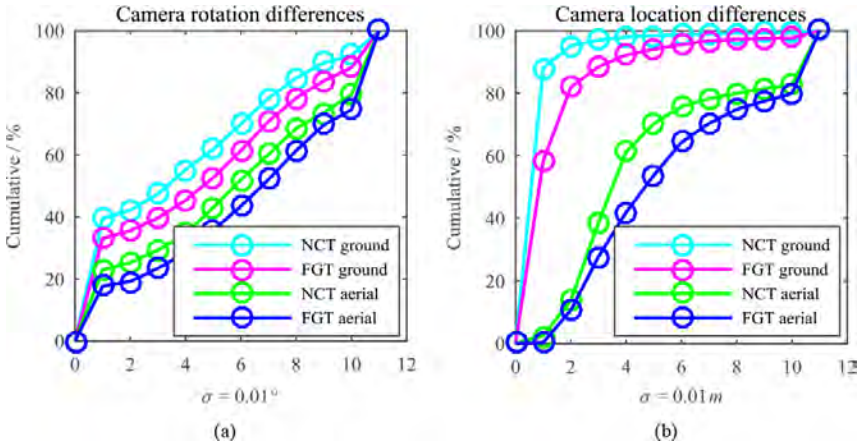

图 3.12: 在 NCT 与 FGT 数据集上经 BA 点云融合前后的相机位姿差异。（a）相机旋转差异。（b）相机位置差异。图中的曲线为累积误差分布曲线，其中 y 轴为（航拍与地面）相机位姿（旋转与位置）差异小于 $n\sigma$ 占所有相机的百分比。

Figure 3.12: Camera pose differences before and after point cloud merging by bundle adjustment on NCT and FGT datasets. (a) Camera rotation differences. (b) Camera location differences. The curves are the cumulative error distributions with the y-axis being the percentage of (ground or aerial) cameras with pose (rotation or location) differences $< n\sigma$.

数据集上差异更大。这说明基于图像的建模方法存在场景漂移的问题，仅通过一个相似变换不足以表征航拍与地面点云之间的变换关系。另外，在 NCT 与 FGT 数据集上，航拍相机的位姿差异均大于地面相机，这是由于航拍相机的位姿估计对观测到的特征点与空间点投影之间的偏差更为敏感。

3.7 拓展：由稀疏到稠密点云

本节对如何根据融合的稀疏点云及相机生成稠密点云进行了简单介绍，这是本章工作的一个直接的拓展。

由于本章中的点云均为由特征点重建得到的稀疏点云，为获取完整且细节丰富的重建重建结果，需要通过 MVS 获取模型的稠密点云，这是基于图像建模的标准步骤。在进行大规模 MVS 时，一个常用做法是先计算每幅图像的深度图然后进行深度图融合。为生成高质量的深度图，需要稀疏点云中每个点的可见性（即哪些相机能够看到该点）用于计算深度图 [75] 与选择邻居图像 [91]。为此，本章方法再一次借助稀疏网格生成跨越航拍与地面图像的可见性信息。

在融合航拍与地面点云之后，本章采用方法 [7] 对融合后点云进行表面重建，得到稀疏网格。并且，本章在此采用3.3.3中的方法，获取每幅航拍与地面图像的可

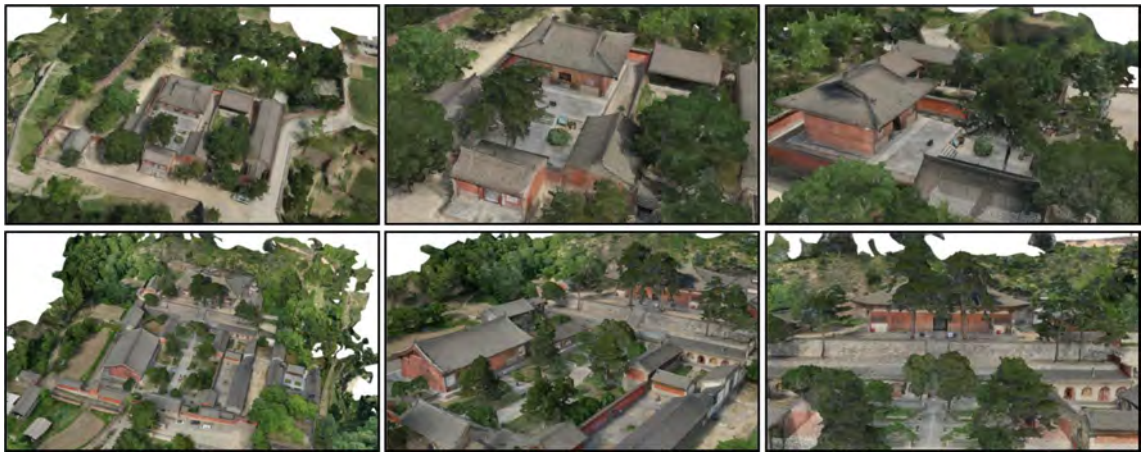

图 3.13: NCT 与 FGT 数据集上的稠密重建结果。第一行：NCT 数据集上的结果。第二行：FGT 数据集上的结果。
Figure 3.13: The dense reconstruction with merged ground and aerial point clouds on the NCT and FGT datasets. First row: results on NCT dataset; second row: results on FGT dataset.

见网格。然后，对于每个面片，可以获取一个可见相机列表，各面片通过其中心表示。上述面片中心以及它们的可见性信息用于选取每幅航拍与地面图像的邻居图像 [75]，然后本章采用方法 [91] 计算每幅图像的深度图并融合成稠密点云。NCT 与 FGT 数据集上的稠密重建结果如图3.13所示。由图3.13可知，通过上述过程重建得到的稠密点云结构完整，细节丰富。

3.8 本章小结

本章在一个统一的框架下实现了航拍与地面图像匹配与点云融合。本章方法通过利用空间连续的网格合成没有孔洞的航拍视角图像用于消除航拍与地面图像在视角与尺度方面的差异。另外，本章还提出了如下两项技术对候选匹配点进行外点过滤：几何一致性检验与几何模型验证。然后，本章方法通过连接航拍与地面特征点轨迹以及 BA 的方式实现了航拍与地面点云的融合。实验结果表明，本章中的航拍与地面图像匹配方法在召回率、精度与效率方面优于其它对比方法；本章中的航拍与地面点云融合方法在精度与效率方面优于其它对比方法。另外，还将本章方法进行了拓展，通过融合后的稀疏点云生成了建筑场景的稠密点云。

第 4 章　融合图像与激光数据的精确完整建模

尽管通过融合航拍与地面图像进行大规模场景三维重建可以获得相比于单一来源图像更为完整的重建结果。然而，基于图像的重建方法依赖环境因素，在场景结构复杂，纹理、光照缺乏的区域效果较差。针对上述问题，本章提出了一种融合图像与激光数据的精确完整建模方法。

4.1　引言

精度与完整性是大规模建筑场景三维重建中的两个关键因素。当前大多数场景重建流程都在关注重建的精度，却对重建完整性不够重视。常见的场景重建方法在结构相对简单的场景下可以取得较好的重建完整性，然而对于较为复杂的建筑场景来说，重建完整性难以保证。为获取大规模复杂场景的完整精确的重建模型（点云或网格），需要对其全局结构与局部细节进行数据采集与重建。当前主要有两种常用的场景重建方式：基于图像的重建 [6, 8–10, 73, 75, 92, 93] 与基于激光扫描的重建 [94–97]。这两类方法在灵活性与精度方面是互补的。

基于图像的重建方法灵活方便，当前最新的图像采集设备具有很好的便携性以及极高的分辨率，十分适合大规模场景的完整采集。然而，现有的基于图像的重建方法对例如光照变化，纹理强度以及结构复杂度等外界因素依赖严重。因此，基于图像的重建结果中错误是很难避免的，尤其是在弱纹理，弱光照以及结构复杂的区域。

基于激光扫描的重建方法精度较高且对一些不利条件更为鲁棒。然而，为了实现对大规模场景的完整覆盖，需要进行多站点激光扫描与对齐。通常，采集的激光点云需要借助人工放置于场景中的人造标志进行粗略对齐，然后利用 ICP 方法进一步实现精细对齐。因此，为实现建筑场景的完整重建，需要进行大量激光扫描与对齐工作，这对目前笨重的扫描设备来说是十分耗时且低效率的。

为通过融合图像与激光数据实现场景的完整重建，一个直接的方法是将图像与激光数据同等对待。具体来说，可以先通过这两类数据分别获取建筑场景模型，然后将它们通过 GCP[42] 或者采用 ICP 算法 [43, 44] 进行对齐。然而，上述做法可行性较差，这是由于通过图像与激光扫描生成的点云在密度、精度以及完整性等方面差异巨大，这样的话会在模型对齐时产生不可避免的误差。另外，需要认真

选取激光扫描位置以确保有足够的重叠区域进行激光扫描数据的自对齐。

本章采用了一种更加有效的数据采集与场景重建方法，该流程可以兼顾数据采集效率以及重建精度与完整性。该方法以图像为主体用于完整覆盖场景，以激光数据为辅助用于针对弱纹理，弱光照以及结构复杂的区域。该方法主要包括三步：图像采集，激光扫描以及图像与激光融合。图像用于完整覆盖场景并通过 SfM 生成稀疏点云。基于 SfM 结果，本章方法自动规划激光扫描位置。最后，图像与激光数据通过一种由粗到细的方式进行融合，以生成精确而完整的场景重建结果。本章方法有如下两个优势：（1）由于在此激光数据仅作为图像的辅助，激光数据之间的重叠以及用于对齐的人造标记都不再是必须的了；（2）通过将激光数据融入到基于图像的重建框架之下，可以反过来提升其精度与完整性。

本章方法的主要由如下三个主要贡献：

- 一个以图像为主，激光数据为辅的新颖的重建流程，该流程兼顾数据采集效率以及重建精度与完整性。
- 一个全自动的激光扫描规划算法，该算法考虑到了场景的纹理强度，结构复杂度以及激光站点的空间分布情况。
- 一个由粗到细图像与激光融合方法，该方法可生成精确、完整的场景重建结果。

4.2 方法概述

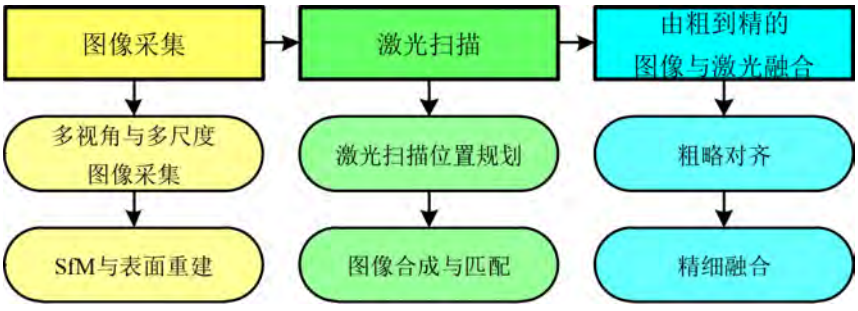

图 4.1: 本章的场景完整重建方法的流程图。该方法主要包括三步：（1）图像采集；（2）激光扫描；（3）由粗到细的图像与激光数据融合。
Figure 4.1: Schematic diagram of the proposed complete scene reconstruction pipeline in this chapter. It mainly contains three steps: (1) image capturing; (2) laser scanning; and (3) coarse-to-fine image and laser scan merging.

本章的场景完整重建方法的流程图如图4.1所示。该方法主要包括三步：（1）图像采集；（2）激光扫描；（3）图像与激光数据融合。在下文中将对各个步骤进行详细介绍。

4.3 图像采集

为了完整覆盖大规模场景，本章方法首先从天上，地上，室内，室外进行了多视角与多尺度的图像采集，然后通过 SfM 获取采集图像对应的稀疏点云。

4.3.1 多视角与多尺度图像采集

表 4.1: 图像采集细节。其中，$a:b:c$ 表示以 a 为起点，c 为终点，b 为步长的一组数。
Table 4.1: Details of image capturing. $a:b:c$ denotes a set of numbers, whose beginning is a, ending is c, and step is b.

	航拍图像	地面图像
采集设备	安装在 Microdrones MD4-1000 上的 Sony NEX-5R	安装在 GigaPan Epic Pro 上的 Canon EOS 5D Mark III
采集模式	5 条航线 1 条用于采集垂直视角图像 4 条用于采集 45° 倾斜视角图像	每站 45 幅图像 俯仰角：$-40°:20°:40°$ 偏航角：$0°:40°:320°$
焦距	$24mm$	$35mm$
图像分辨率	$4912px \times 3264px$	$5760px \times 3840px$

为保证对建筑场景的完整覆盖，本章通过如下两种方式进行图像采集：（1）有着较好连接性的近距离地面图像用于覆盖室内与室外场景；（2）大尺度航拍图像用于整个场景与建筑屋顶的采集。一些图像采集的具体细节如表4.1及图4.2所示。其中，（室内与室外）地面图像以站的形式进行采集，这样的话可以方便规划图像采集位置，提高图像采集效率。另外，为了从地面视角较为适当地覆盖室内与室外场景，地面图像采集站在整个场景中分布较为均匀。具体来说，本章方法对感兴趣场景地平面进行粗略网格划分并将网格中心用作图像采集位置，并将网格边长设为 $3m$。在进行图像采集时，本章方法在地面上对上述位置进行标记，这些位置在后续激光扫描过程中会再次用到。

图 4.2: 本章中的数据采集设备。第一列：安装在 GigaPan Epic Pro 上的 Canon EOS 5D Mark III，用于地面图像采集；第二列：安装在 Microdrones MD4-1000 上的 Sony NEX-5R，用于航拍图像采集；第三列：Leica ScanStation P30 Scanner，用于地面激光数据采集。
Figure 4.2: Data collection equipments in the experiments of this chapter. First column: a Canon EOS 5D Mark III mounted on a GigaPan Epic Pro for ground image capturing; Second column: a Sony NEX-5R mounted on a Microdrones MD4-1000 for aerial image capturing; Third column: a Leica ScanStation P30 Scanner for terrestrial laser scanning.

4.3.2 SfM 与表面重建

接下来，本章方法将采集的图像通过 SfM[73] 进行相机位姿标定并生成稀疏点云。为了将所有采集到的（室内，室外与航拍）图像融合到统一的 SfM 过程中，需要航拍与地面图像匹配点以及室内与室外图像匹配点。然而，获取上述两类匹配点均不容易。这是由于：（1）航拍与地面图像在视角与尺度方面差异巨大；（2）室内与室外图像公共可见区域十分有限。在此，本章采用上一章的方法获取航拍与地面图像匹配点，图4.3给出了一对航拍与地面图像匹配结果示例。另外，最近有一种方法 [98] 通过识别并对齐窗户的方式实现室内场景与室外场景的对齐。然而，这种方法并不适用于所有建筑类型。本章通过匹配位于门附近的室内与室外图像实现场景的融合，图4.4给出了一对室内与室外图像匹配结果示例。由图4.4可知，获取的室内与室外图像匹配点足够进行室内与室外的场景融合。由于基于图像的重建方法具有尺度不确定性，为了能够规划激光扫描位置以及融合 SfM 与激光点云，需要恢复融合的（室内，室外与航拍）SfM 点云的真实尺度。本章中，该尺度通过相机内置 GPS 粗略恢复。在此之后，本章采用方法 [7] 对融合点云进行表面重建，获取场景的三维网格。该三维网格将用于后续的激光扫描位置规划。

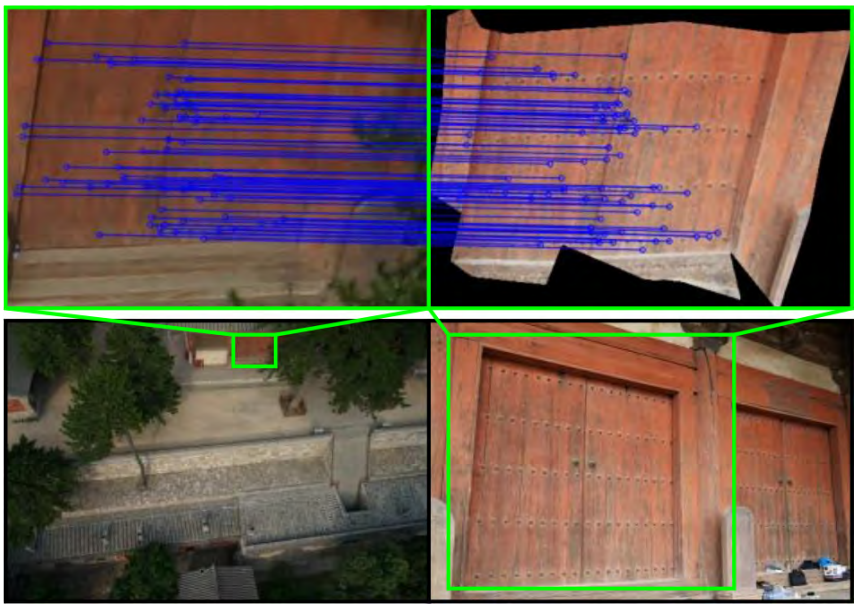

图 4.3: 一对航拍（左图）与地面（右图）图像匹配结果示例。第一行：剪裁的航拍图像与合成的航拍视角图像之间的匹配结果，其中蓝色线段表示匹配点。第二行：视角与尺度差异明显的航拍与地面图像。

Figure 4.3: An example of ground-to-aerial image feature matching result. First row: point matches between the cropped aerial image (left) and synthetic aerial-view image (right), where the blue segments link the point matches; second row: original aerial and ground image pair with large viewpoint and scale differences.

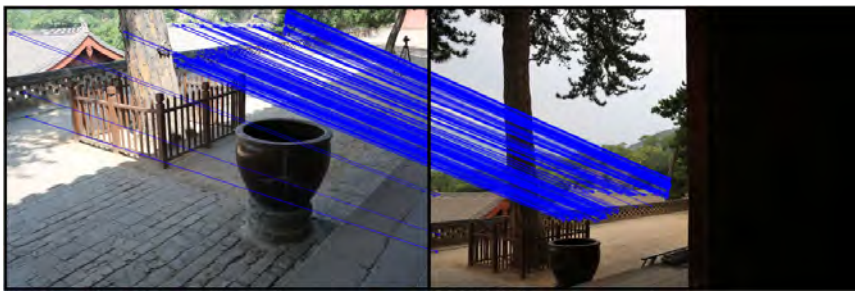

图 4.4: 一对室内（右图）与室外（左图）图像匹配结果示例，其中蓝色线段表示匹配点。

Figure 4.4: An example of outdoor(left)-to-indoor(right) image feature matching result, where the blue segments link the point matches.

4.4 激光扫描

基于 SfM 结果，本章方法通过考虑如下三个因素实现激光扫描位置的自动规划：纹理强度，结构复杂度以及激光站点的空间分布情况。然后，为融合图像与激光数据，本章方法从激光数据中合成航拍与地面视角图像并与采集到的图像进行匹配。

4.4.1 激光扫描位置规划

为了通过尽量少的激光扫描数据完整覆盖场景，目前有一些方法 [99–102] 针对最优地面激光扫描网络设计问题提出了解决方案。这些方案基于二维建筑图纸 [99, 101, 102] 或者三维建筑模型 [100]。在进行最优化求解时，需要考虑各种因素。例如距离与入射角 [99]，足够的重叠与表面拓扑 [100] 以及多尺度与分层视点规划 [102]。然而，本章引入激光的目的是为了获取更加精确完整的场景重建结果且激光仅用作图像的辅助。因此，本章方法在规划激光扫描位置时考虑如下三个因素：（1）场景的纹理强度，（1）场景的结构复杂度以及（3）激光站点的空间分布情况。前两个因素意味着弱纹理与结构复杂的区域需要优先进行激光扫描以对其进行补充。第三个因素表示激光扫描位置应在场景中较为均匀地分布且彼此之间不应重叠过多。根据上述三个因素，本章提出了一个激光扫描位置的自动规划方法。

为规划激光扫描位置，本章方法首先获取一些候选激光扫描位置。这样的话，激光扫描位置规划问题就转变成了一个 0-1 整数规划问题：从候选扫描位置中选取一些作为真实扫描位置（标记为 1），并舍弃其它位置（标记为 0）。候选激光扫描位置可以通过如下方式较为便捷地获取。首先检测场景中的地平面并将其按照栅格进行划分。然后将每个栅格的中心作为候选激光扫描位置 [99, 101, 102]。然而，本章方法将地面图像采集位置作为候选激光扫描位置，这样做有如下两个优势：（1）图像采集位置是为了适当地覆盖场景认真选取确定的，因此这些位置的子集也适用于进行激光扫描；（2）在后续的图像与激光数据融合过程中，需要获取采集的地面图像与由激光数据合成的地面视角图像之间的匹配点，而在图像采集位置进行激光扫描有助于进行图像的合成与匹配。

在获取了候选激光扫描位置后，本章方法借助由稀疏点云生成的三维网格确定真实激光扫描位置。具体来说，本章方法在每个候选扫描位置均匀投射 n_r 条射线 [103]，其中 $n_r = 1000$。由第 i 个候选扫描位置发出的射线与三维网格共有 n_i 个面片相交，相交面片通过 CGAL 库获取，记为：

$$\mathcal{F}_i = \{f_{i,m}\}(i = 1, 2, \ldots, N_p; m = 1, 2, \ldots, n_i) \tag{4.1}$$

其中 N_p 为候选激光扫描位置数量，$f_{i,m}$ 为第 i 个位置与射线相交的第 m 个面片。这些通过射线相交得到的面片用于衡量候选扫描位置周围场景的纹理强度以及结构复杂度。具体来说，对于每个面片 $f_{i,m}$，本章方法获取距它小于 r_f 的所有面片，

记为其邻域面片，其中 $r_f = 0.1m$。在这里面片之间的距离定义为面片中心之间的欧氏距离。然后，本章方法将 $f_{i,m}$ 与其邻域面片面积加和，记做 $a_{i,m}$。$a_{i,m}$ 的值用于衡量 $f_{i,m}$ 附近场景的纹理强度以及结构复杂度：$a_{i,m}$ 的值越大，$f_{i,m}$ 附近场景纹理越弱，结构越复杂。本章方法进而通过如下表达式衡量第 i 个候选激光扫描位置的纹理强度与结构复杂度：

$$A_i = \frac{\sum_{m=1}^{n_i} a_{i,m}}{n_i} \tag{4.2}$$

另外，本章方法采用相交面片集合的 IoU 来衡量第 i 个与第 j 个候选扫描位置之间的重叠情况，并将其记为：

$$IoU_{i,j} = \frac{\mathcal{F}_i \bigcap \mathcal{F}_j}{\mathcal{F}_i \bigcup \mathcal{F}_j} \tag{4.3}$$

由于希望激光扫描位置分布更为均匀，本章方法选取互相之间 IoU 更小的一些候选扫描位置。因此，本章方法将激光扫描位置规划问题阐述如下：

$$\begin{aligned} &\max_{x_i} \frac{\sum_{i=1}^{N_p} A_i x_i}{\sum_{i=1}^{N_p} \sum_{j=i+1}^{N_p} x_i x_j IoU_{i,j}}, \\ &s.t. \bigcup_{i=1}^{N_p} x_i \mathcal{F}_i \Big/ \bigcup_{i=1}^{N_p} \mathcal{F}_i < t_c \end{aligned} \tag{4.4}$$

其中，$x_i = 0, 1, (i = 1, 2, \ldots, N_p)$ 为优化变量，$x_i = 1$ 表示本章方法选取了第 i 个候选扫描位置，否则的话 $x_i = 0$；t_c 为一个边界阈值，用于限制激光数据对场景的覆盖程度，本章中 $t_c = 0.25$。

然而，式4.4定义的问题为一个 0-1 整数规划问题，该问题为 NP 问题。本章采用一个贪婪算法对该问题近似求解。该算法每次选取一个候选扫描位置作为最终的真实扫描位置，算法细节描述如下。

本章方法首先将有着最大 A_i 的候选扫描位置选为第一个真实扫描位置：

$$i_1^* = \arg\max A_i \tag{4.5}$$

上式表示第 i_1^* 个候选扫描位置为第一个选取的真实扫描位置。假设在进行完 N_s 次扫描位置选取后，选取的 N_s 个扫描位置在候选扫描位置集合中的序号记为

算法 2: 激光扫描位置规划算法

Input :
定义于式4.2中的 A_i 以及定义于式4.3中的 $IoU_{i,j}$。
Output:
选取的候选激光扫描位置序号。

初始化：
根据式4.5选取第一个激光扫描位置，$N_s \leftarrow 1$。
迭代选取：
repeat
　根据式4.7选取一个激光扫描位置，$N_s \leftarrow N_s + 1$。
until 满足定义于式4.8中的条件;

$\{i_m^* | m = 1, 2, \ldots, N_s\}$，而剩余的未选取的序号记为 $\{i_n^\# | n = 1, 2, \ldots, N_p - N_s\}$。这意味着：

$$\begin{aligned} x_{i_1^*} = x_{i_2^*} = \cdots = x_{i_{N_s}^*} = 1 \\ x_{i_1^\#} = x_{i_2^\#} = \cdots = x_{i_{N_p-N_s}^\#} = 0 \end{aligned} \tag{4.6}$$

然后，在第 $N_s + 1$ 次选取中，本章方法通过如下优化从 $\{i_n^\#\}$ 中选取：

$$n^* = \arg\max \frac{\sum_{m=1}^{N_s} A_{i_m^*} + A_{i_n^\#}}{\sum_{m=1}^{N_s} \sum_{k=m+1}^{N_s} IoU_{i_m^*, i_k^*} + \sum_{m=1}^{N_s} \sum_{n=1}^{N_p - N_s} IoU_{i_m^*, i_n^\#}} \tag{4.7}$$

选取的扫描位置在集合 $\{i_n^\#\}$ 的序号为 n^*，这意味着 $i_{N_s+1}^* = i_{n^*}^\#$。随着选取过程的进行，本章方法选出来了越来越多的真实扫描位置，整个选取过程通过式4.4中的截止条件停止：

$$\bigcup_{m=1}^{N_s} \mathcal{F}_{i_m^*} \Big/ \bigcup_{i=1}^{N_p} \mathcal{F}_i < t_c \tag{4.8}$$

本章的激光扫描位置自动规划算法总结于算法2中。需要注意的是，由于本章方法在规划的扫描位置进行激光扫描，生成的激光点云与 SfM 点云均处于地理坐标系下，简化了后续的图像合成与匹配过程。

4.4.2 图像合成与匹配

在激光扫描位置规划完成后，本章方法在规划的位置进行地面激光扫描，采用的扫描设备为 Leica ScanStation P30 Scanner（见图4.2）。像其它当前先进的激光

扫描仪一样，P30 可以获取大量、精确且带有 RGB 信息的三维空间点。为了融合图像与激光数据，本章方法通过激光点合成图像并与采集的图像进行匹配。本章不仅是像文献 [12] 中一样合成地面视角图像，还通过激光点云对航拍视角图像进行了合成。通过匹配航拍与合成图像，本章方法可以获取更多的用于后续图像与激光数据融合的约束。

4.4.2.1 地面视角图像合成

图 4.5: （a）用于地面视角图像合成的虚拟立方体示意图，其中蓝色棱锥表示其中的一个虚拟相机。（b）一个地面视角和成图像示例。（c）图像（b）的深度图。
Figure 4.5: (a) Schematic diagram of the virtual cube for ground-view image synthesis, where the blue pyramid denotes one of the virtual cameras. (b) An example of ground-view synthetic image. (c) Depth map of (b).

对于每个激光扫描位置获取的点云，本章方法将其投影至一个虚拟立方体的六个面上获取六幅合成图像。虚拟立方体的中心与激光扫描原点重合，如图4.5a 所示。合成图像像素的 RGB 值即为投影至该像素的激光点的 RGB 值（见图4.5b）。虚拟立方体的六个面以及中心构成了六个朝向互相垂直的相机。该模型可以视作一个广义相机模型 [104, 105]。地面视角合成图像的宽高均设为采集的地面图像的高度，本章中为 3840px。

4.4.2.2 航拍视角图像合成

在此，本章方法采用2.4节中的方法选取合适的航拍图像并由激光点云合成选取的航拍视角图像。对于每个激光扫描位置，本章方法首先选取五幅航拍视角图像。这五幅图像可以较为合理地覆盖采集到的激光数据且分布较为均匀，然后将每幅图像的可见激光点通过相机的内外参数投影至该图像上以合成航拍视角图像。

由于此处的航拍与地面视角图像均通过点云投影的方式合成，本章方法对合成图像进行最近邻插值以填补一些不可避免的缺失像素。另外，由于后续过程中需要用到合成图像像素与激光点云之间的二维三维对应关系，每幅合成图像均生

成了对应的深度图，如图4.5c 所示。

图 4.6: 一对合成（左图）与地面（右图）图像匹配结果示例，其中蓝色线段表示匹配点。
Figure 4.6: An example of synthetic-to-ground image feature matching result, where the blue segments link the point matches.

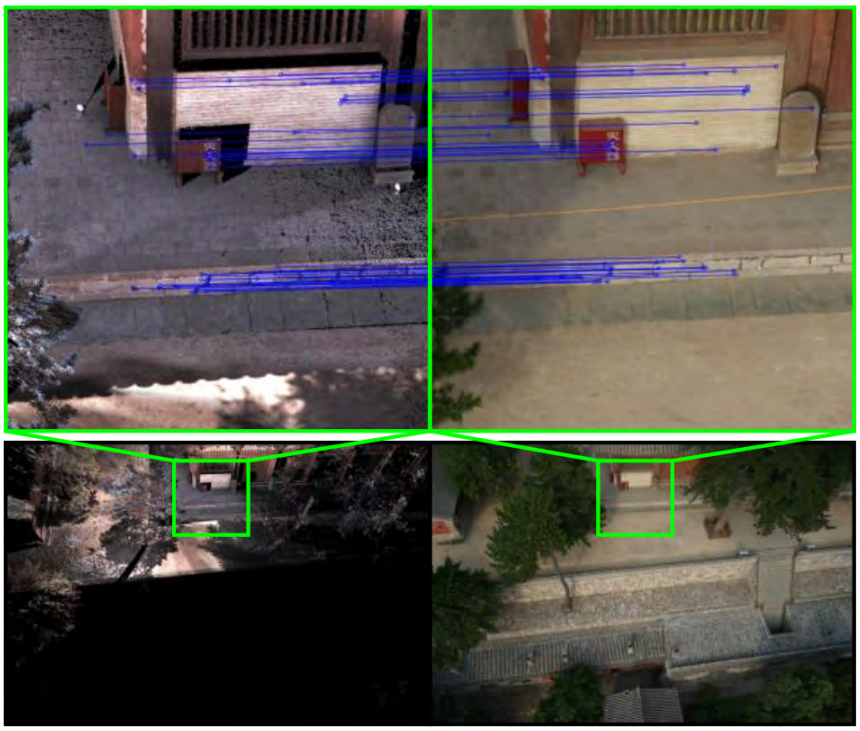

图 4.7: 一对合成（左图）与航拍（右图）图像匹配结果示例。第一行：第二行中绿色矩形区域的放大图像，为更好的展示由蓝色线段表示的匹配点。第二行：原始合成与航拍图像对。
Figure 4.7: An example of synthetic-to-aerial image feature matching result. First row: enlarged synthetic-to-aerial image pair of the green rectangles in the second row to illustrate the feature point matches, which are denoted by the blue segments; second row: original synthetic and aerial image pair.

接下来，本章方法对合成与采集图像进行 SIFT 特征匹配。地面视角合成图像与距离和朝向均比较接近的地面采集图像进行匹配，本章中距离小于 $5m$，朝向夹

角小于 45°。航拍视角合成图像与图像合成时的目标航拍图像进行匹配。另外，由于合成图像边缘处的深度是不可靠的，在进行图像匹配之前，本章方法舍弃掉了位于合成图像 Canny 边缘 [106] 附近的特征点。合成与地面图像以及合成与航拍图像的匹配结果示例分别如图4.6以及图4.7所示。

4.5 由粗到精的图像与激光融合

在进行图像与激光数据融合时，首先，激光数据逐一地粗略对齐至 SfM 点云；然后，借助跨属性的匹配点，通过广义 BA 实现图像与激光数据的精细融合。

4.5.1 粗略对齐

本章方法将4.4节中获取激光扫描数据通过如下方式逐一地粗略对齐至 SfM 点云。在此，将第 i 个扫描位置获取的激光点云与 SfM 点云之间的相似变换记做 $\{s_i, \boldsymbol{R}_i, \boldsymbol{t}_i\}$。用于估计该相似变换的三维对应点由4.4.2节获取的合成与地面图像以及合成与航拍图像二维匹配点转换得到。本章方法仍通过 RANSAC 对该相似变换进行估计，在 RANSAC 过程中，本章方法进行 100 次随机抽样且将距离阈值设为 $0.1m$。在估计完相似变换后，本章方法将第 i 个扫描位置获取的激光点云与 SfM 点云之间的合成与地面以及合成与航拍三维对应点内点分别记为 $\{\boldsymbol{X}_{i,m}^{GL} \leftrightarrow \boldsymbol{X}_{i,m}^{GI}\}$ 与 $\{\boldsymbol{X}_{i,n}^{AL} \leftrightarrow \boldsymbol{X}_{i,n}^{AI}\}$。

4.5.2 精细融合

在将激光数据粗略对齐至 SfM 点云之后，本章方法通过广义 BA 联合优化（室内、室外与航拍）相机姿态，融合的 SfM 点云以及 SfM 点云与激光点云之间的相似变换，以精细融合图像与激光数据。进行上述进一步优化的原因有如下两部分：（1）SfM 点云可能不够精确，甚至存在场景漂移现象，这种情况在大规模场景中尤为明显；（2）航拍与地面 SfM 点云可能并未精确地融合在一起。通过将 SfM 结果（相机位姿与空间点坐标）与激光数据对齐结果（相似变换）融合到一个全局优化中，上述两问题的精度均可提升。本章方法将这里的 BA 过程称为广义 BA 的原因是，在此 BA 过程中通过最小化三维到二维的重投影误差以及三维到三维的空间误差以同时优化相机位姿以及用于激光数据对齐的相似变换。

采集图像的相机姿态，融合的 SfM 点云以及用于激光数据对齐的相似变换通

过如下方式同时优化：

$$\begin{aligned} &\min_{\theta} \sum_{j} \sum_{k} E_R(j,k), \\ &s.t. \sum_{i} \left(\sum_{m} E_S^G(i,m) + \sum_{n} E_S^A(i,n) \right) = 0 \end{aligned} \tag{4.9}$$

其中，$\theta = \{\boldsymbol{R}_j, \boldsymbol{t}_j, \boldsymbol{X}_k, s, \boldsymbol{R}_i, \boldsymbol{t}_i\}$ 为待优化参数。$\boldsymbol{R}_j$ 与 $\boldsymbol{t}_j$ 为第 j 个相机的旋转矩阵与平移向量；$\boldsymbol{X}_k$ 为融合的 SfM 点云中的第 k 个点；s 为激光对齐的全局尺度；$\boldsymbol{R}_i$ 与 $\boldsymbol{t}_i$ 为第 i 个扫描位置获取的激光点云对齐的旋转矩阵与平移向量。在式4.9中优化全局尺度 s 的原因在于：在4.3.2节中通过相机内置 GPS 恢复的 SfM 点云的尺度精度较低。为实现图像与激光数据的高精度融合，需要精确求取 SfM 点云与激光点云之间的尺度比，以获取 SfM 点云的真实尺度。

式4.9中的重投影误差项 $E_R(j,k)$ 定义如下：

$$E_R(j,k) = \| \boldsymbol{x}_{j,k} - \gamma(\boldsymbol{K}_j, \boldsymbol{R}_j, \boldsymbol{t}_j, \boldsymbol{X}_k) \|^2_{\Sigma^{-1}_{j,k}} \tag{4.10}$$

其中 $\boldsymbol{x}_{j,k}$ 为点 $\boldsymbol{X}_k$ 在第 j 幅图像中观测到的投影点；$\boldsymbol{K}_j$ 为第 j 个相机的内参矩阵，由于 $\boldsymbol{K}_j$ 在4.3节中已经精确标定，此处优化时其值保持不变；$\gamma(\cdot)$ 为投影方程；$\Sigma_{j,k}$ 为 $\boldsymbol{x}_{j,k}$ 的协方差矩阵，其值与 $\boldsymbol{x}_{j,k}$ 局部特征的尺度相关。

式4.9中的地面空间误差项 $E_S^G(i,m)$ 与航拍空间误差项 $E_S^A(i,m)$ 分别定义为：

$$\begin{aligned} E_S^G(i,m) &= \| s\boldsymbol{R}_i \boldsymbol{X}_{i,m}^{GL} + \boldsymbol{t}_i - \boldsymbol{X}_{i,m}^{GI} \|^2_{\Sigma^{-1}_{i,m}} \\ E_S^A(i,n) &= \| s\boldsymbol{R}_i \boldsymbol{X}_{i,n}^{AL} + \boldsymbol{t}_i - \boldsymbol{X}_{i,n}^{AI} \|^2_{\Sigma^{-1}_{i,n}} \end{aligned} \tag{4.11}$$

其中，$\Sigma_{i,m}$ 与 $\Sigma_{i,n}$ 分别为 $\boldsymbol{X}_{i,m}^{GL}$ 与 $\boldsymbol{X}_{i,n}^{AL}$ 的协方差矩阵，其值与激光点到扫描原点距离相关。

在式4.10与式4.11中引入马氏距离是为了消除重投影误差项与空间误差项在尺度与噪声程度上的不均衡。由上述定义可知，式4.9是在限定（理想状况下）没有定义于式4.11中的（航拍与地面）激光与 SfM 点空间误差时最小化定义于式4.10中的 SfM 点云的重投影误差。因此，式4.9定义的优化问题为带约束的优化问题，本

章中通过拉格朗日乘子法进行求解：

$$\min_{\theta}\left(\sum_{j}\sum_{k}\rho\big(E_R(j,k)\big)+\omega\sum_{i}\left(\sum_{m}\rho\big(E_S^G(i,m)\big)+\sum_{n}\rho\big(E_S^A(i,n)\big)\right)\right) \tag{4.12}$$

其中，$\rho(\cdot)$ 为 Huber 损失函数，用于应对不可避免的误匹配与噪声情况；ω 为拉格朗日乘子，用于控制定于式4.10与式4.11中约束的权重。本章方法采用 Ceres Solver 对上式定义的问题进行求解。注意，当 $\omega \to 0$ 时，式4.12中的优化问题主要优化的是（三维到二维的）重投影误差，趋近于标准的 BA 问题；而当 $\omega \to \infty$ 时，式4.12中的优化问题主要是优化（三维到三维的）空间误差，趋近于激光数据对齐问题。本章方法在后续的实验部分描述并评测了一个可以自适应设置参数 ω 的方法。

4.6 实验结果

表 4.2: 用于方法测评的数据集元数据。
Table 4.2: Meta-data of the datasets for method evaluation.

数据集	NCT	FGT
覆盖面积	$3100m^2$	$34000m^2$
地面室外图像数量	2790	6975
地面室内图像数量	855	2475
地面室外图像采集时间	$124min$	$310min$
地面室内图像采集时间	$57min$	$165min$
室外室内图像数量比	3.26	2.82
航拍图像数量	772	1596
室外激光扫描站数	6	19
室内激光扫描站数	5	14
室外激光扫描时间	$180min$	$570min$
室内激光扫描时间	$200min$	$560min$
室外室内激光站数比	1.20	1.36

本节在两组中国古代建筑数据集，南禅寺（NCT）与佛光寺（FGT）上对本章方法进行了测评。上述两个寺庙中大殿的室内场景均有着复杂的结构且光线较弱。因此，对于本章研究的问题来说，它们是合适的研究载体。上述两数据集的元数据列于表4.2中。需要注意的是，对于地面（室外、室内）图像，本章方法在每个图像采集位置采集了 45 幅图像（见表4.1），这意味着对于 NCT 室外、室内场景，FGT

室外、室内场景，分别有 62(2790/45)，19(855/45)，155(6975/45) 以及 55(2475/45) 图像采集位置。另外，本节将地面图像以及激光数据采集时间列于表4.2中。地面室外、室内图像采集平均长为 $2min$、$3min$ 每站；室外、室内激光扫描平均时长为 $30min$、$40min$ 每站。由于室内场景光照较弱，曝光时间较长，因此在室内场景中数据采集时间也相应更长。

4.6.1 图像采集结果

本节采用4.3节中介绍的流程进行采集图像与生成 SfM 点云。在 NCT 与 FGT 数据集中采集的图像示例以及重建得到的 SfM 点云如图4.8所示。由图4.8可知，航拍与（室内、室外）地面 SfM 点云较好地融合在了一起。然而，在 NCT 与 FGT 数据中的弱纹理，弱光照以及结构复杂区域的重建得到的点的数量过少。因此，进行激光扫描以获取更加精确、完整的建筑场景模型是十分必要的。

4.6.2 激光扫描结果

然后，本节采用4.4节中的方法规划激光扫描位置并进行激光扫描。如表4.2所示，在 NCT 与 FGT 数据上，室内室外图像数量比的值均大于室内室外激光站数比的值。另外，由于候选激光扫描位置，即地面图像采集站的位置，是等间距分布的，更小的室内室外激光站数比意味着通过本章的激光扫描位置规划算法，室内激光扫描站更加密集。这是由于相对于室外场景，室内场景结构更为复杂，纹理、光照更弱。在 NCT 与 FGT 数据集上的激光扫描位置规划结果如图4.9所示。由图4.9可知，本章方法规划的激光扫描位置较为均匀、稀疏地分布于整个建筑场景中。

另外，为了验证本章激光扫描位置规划方法，本节分别从 NCT 与 FGT 数据中选取了一个规划激光扫描位置，对其附近同一区域的激光点云与 SfM 点云进行比较，如图4.10所示。该区域在图4.9第一列由蓝色矩形表示，而扫描位置在图4.9第三列由蓝色圆表示。本节在 NCT 数据上选取了一个室内激光扫描位置，在 FGT 数据上选取的是一个室外激光扫描位置。由图4.10可知，标示区域仅有较少的 SfM 点，这是由于这些区域纹理较弱（如墙面）或者结构过于复杂（如室内的彩塑以及室外的斗拱）。因此，通过本章的激光扫描位置规划方法，激光扫描可以有效地覆盖建筑场景中的弱纹理与复杂结构区域以获取更为精确、完整的建筑场景模型。

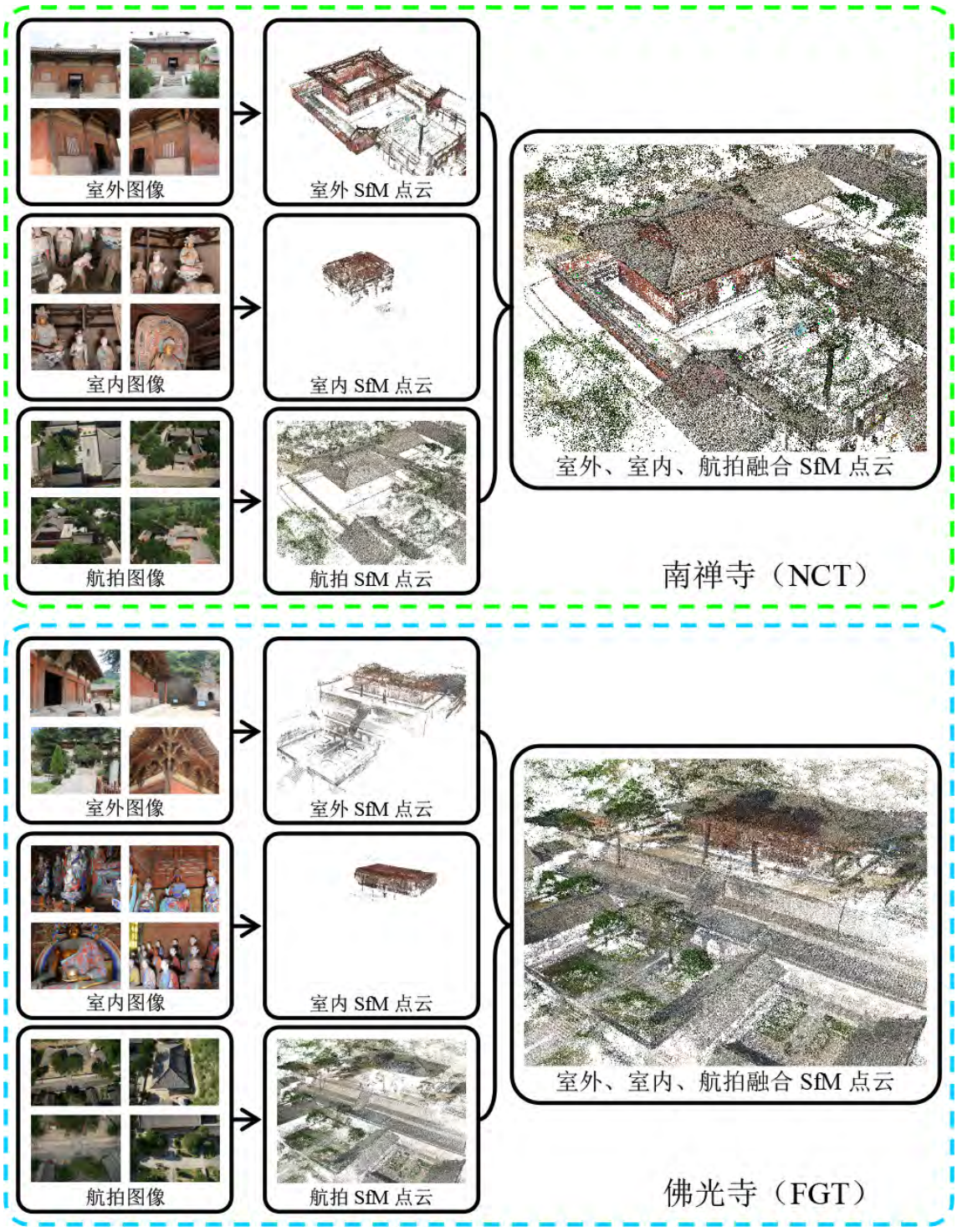

图 4.8: NCT（上图）与 FGT（下图）数据的示例图像以及融合的 SfM 点云。
Figure 4.8: Examples of captured images and merged SfM points of NCT (top) and FGT (bottom).

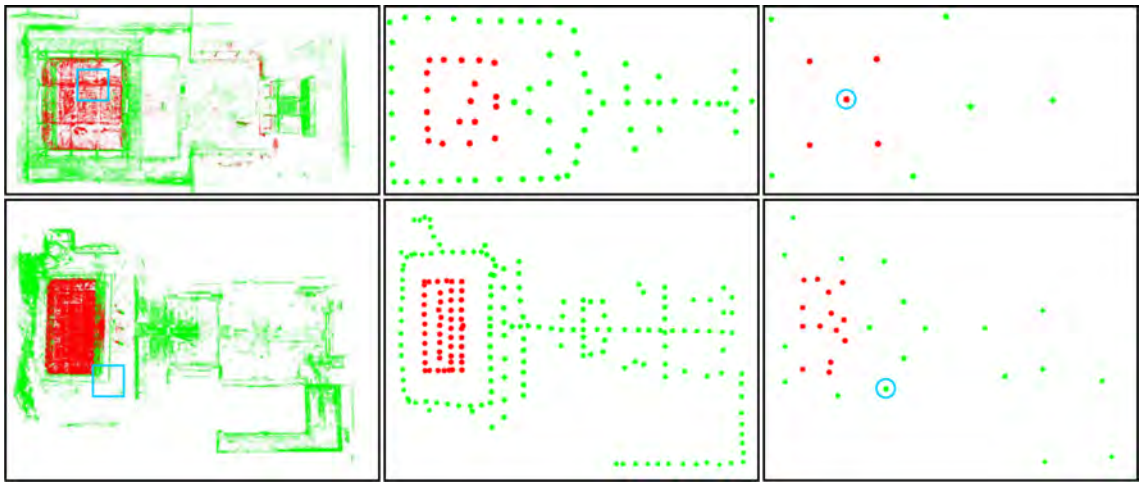

图 4.9: NCT 与 FGT 数据集上的激光扫描位置规划结果。第一行：NCT 数据集上的结果；第二行：在 FGT 数据集上的结果。第一列：融合的地面室内（红色）与室外（绿色）SfM 点云；第二列：室内（红色）与室外（绿色）候选激光扫描位置；第三列：室内（红色）与室外（绿色）规划激光扫描位置。
Figure 4.9: Laser scanning location planning results on NCT and FGT. First row: result on NCT; second row: result on FGT. First column: merged ground outdoor (green) and indoor (red) SfM points; second column: outdoor (green) and indoor (red) potential laser scanning locations; third column: outdoor (green) and indoor (red) planned laser scanning locations.

4.6.3 图像与激光融合结果

接下来，本节采用4.4节中的方法对图像与激光数据进行由粗到细的融合。定性与定量结果均在下文中给出。

4.6.3.1 定性结果

NCT 与 FGT 数据集上的定性结果如图4.11所示。为了更好的视觉效果，本节对融合后的 SfM 与（降采样的）激光点云通过方法 [7] 进行表面重建。通过图4.11中的远景图可知图像与激光数据较好地融合在了一起。而通过图4.11中的近景图可知本章方法即使是在弱纹理与结构复杂区域也能获得精确、完整的场景重建结果。因此，上述定性结果验证了本章中的图像与激光数据融合方法的有效性。

4.6.3.2 定量结果

本节对图像与激光数据融合方法进行了定量测评。首先，本节引入了一种用于定量测评融合精度的评价指标。然后，基于该指标，本节对图像与激光融合时参数 ω 的设定进行了评估。最后，本节将本章融合算法与两种方法 [12, 13] 进行了比较。

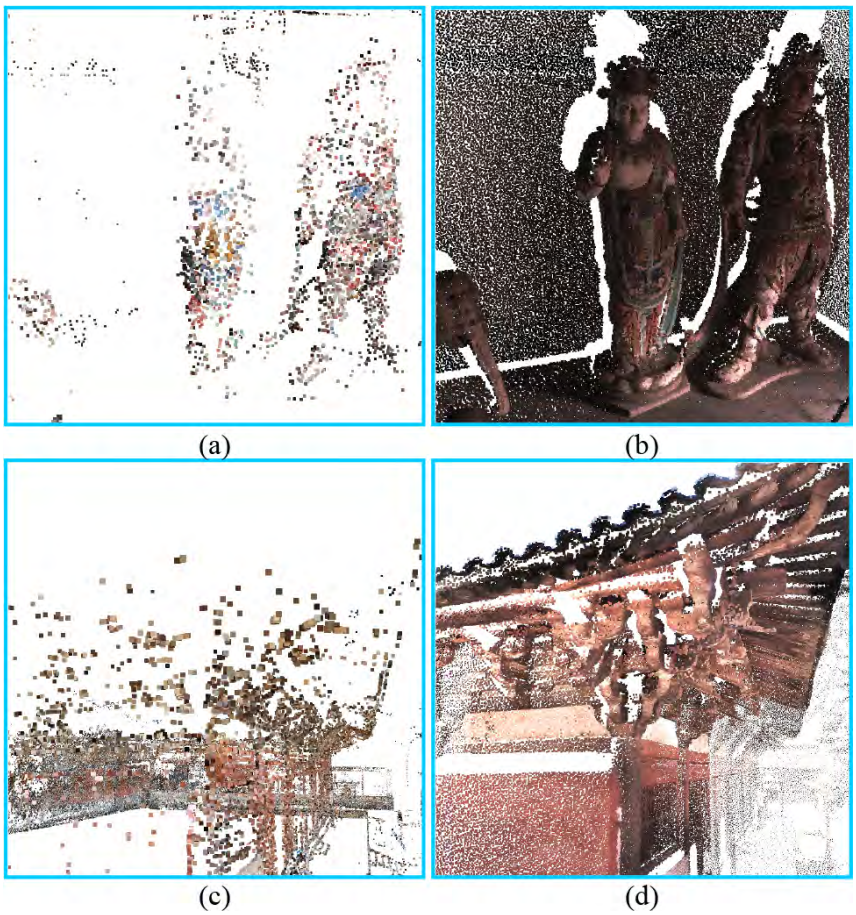

图 4.10: （a）-（b）图4.9左上角蓝色矩形标示区域的 SfM 点云（a）与激光点云（b）。（c）-（d）图4.9左下角蓝色矩形标示区域的 SfM 点云（c）与激光点云（d）。
Figure 4.10: (a)-(b) SfM and laser points of the region marked by blue rectangle in top left corner of Fig. 4.9 (c)-(d) SfM and laser points of the region marked by the blue rectangle in bottom left corner of Fig. 4.9.

评测指标：由于定义一个定量评测融合精度的准确度量十分困难，在此本节采用一种近似测量方法用于定量评测。具体来说，本节首先在 SfM 与激光点云上人工获取一些空间对应点。对于 NCT 与 FGT 数据集，均获取 40 对相对均匀地分布在场景中的参考点，室内场景 20 对，室外场景 20 对。在图像与激光数据融合后，每对参考点理想状态下均应重合。在此将每对参考点之间距离的均方根值用作图像与激光数据融合精度的近似评测指标，平均值越低融合精度越高。

参数设定：尽管通过在式4.10与式4.11引入马氏距离以消除重投影误差项与空间误差项之间在尺度与噪声程度上的不均衡，还存在着另外一种没有考虑到的不均衡因素。该因素为采集图像本身之间的匹配点数量与合成图像与采集图像之间的匹配点数量的差异，会较大程度上影响图像与激光数据融合精度。本章通过引入拉格朗日乘子（式4.12中的 ω）的方式解决上述问题。在此，本节提出了一种自

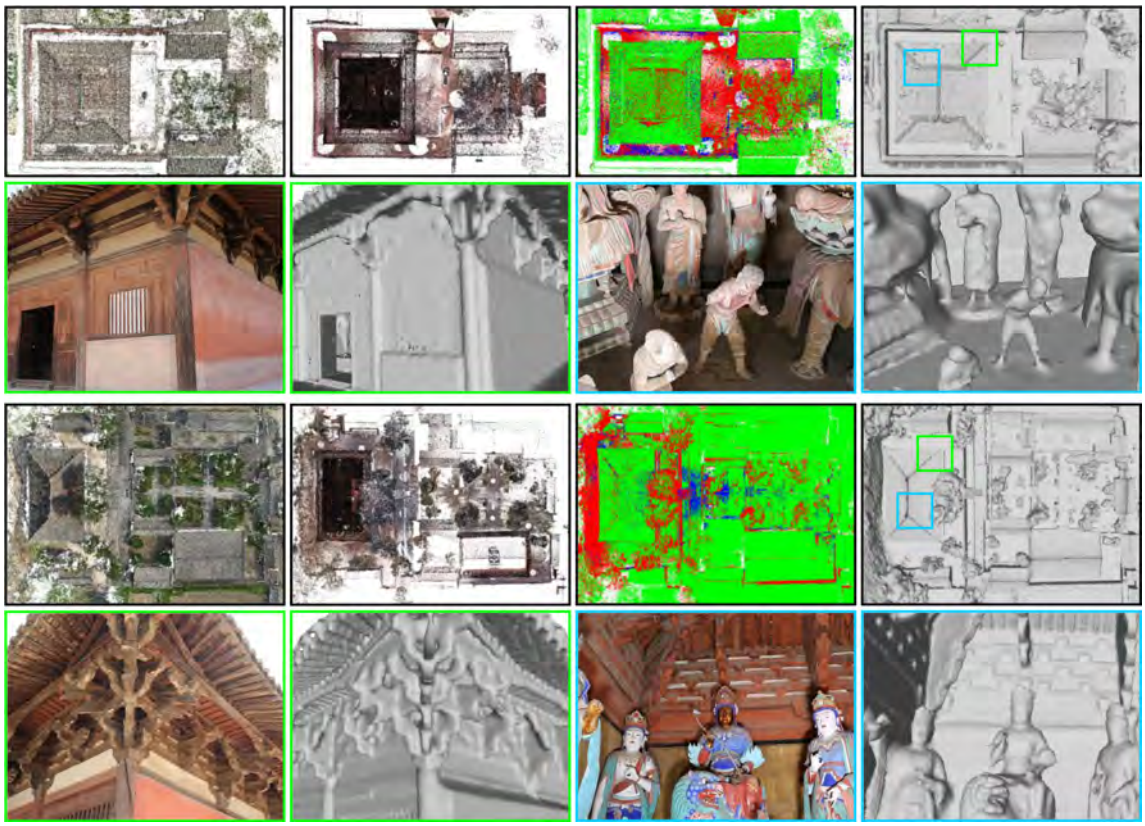

图 4.11: NCT 与 FGT 数据集上的图像与激光数据融合定性结果。第一行：NCT 数据结果远景图，从左到右依次为（室内、室外与航拍）SfM 点云，（室内、室外）激光点云，融合的 SfM 与激光点云（其中红色为激光点云，绿色为航拍 SfM 点云，蓝色为地面 SfM 点云），由融合点云生成的网格。第二行：NCT 数据示例图像以及与示例图像视角类似的网格近景图，左边两个为图中右上角绿色矩形对应的室外区域，右边两个为图中右上角蓝色矩形对应的室内区域。第三行以及第四行：与第一行以及第二行类似的在 FGT 数据集上的结果。

Figure 4.11: Qualitative results of image and laser scan merging on NCT and FGT. First row: long-shots of NCT; from left to right: (outdoor-indoor-aerial) SfM points, (outdoor-indoor) laser points, merged SfM and laser points (red for laser points, green for aerial SfM points, and blue for ground SfM points), surface mesh generated from merged points. Second row: image examples and close-ups of the surface mesh with similar viewpoints on NCT; left two: an outdoor region of the green square at the top right corner of the figure; right two: an indoor region of the blue square at the top right corner of the figure. Third and fourth rows: the results on FGT similar to those of the first and second rows.

适应确定 ω 取值的方法。

根据4.5节，式4.12中的优化问题同时优化相机位姿，SfM 点空间位置以及用于激光数据对齐的相似变换。直观上来讲，当重投影误差代价近似等于空间误差代价时，即：

$$\sum_j \sum_k \rho\big(E_R(j,k)\big) \approx \omega \sum_i \Big(\sum_m \rho\big(E_S^G(i,m)\big) + \sum_n \rho\big(E_S^A(i,n)\big) \Big) \tag{4.13}$$

表 4.3: 在 NCT 与 FGT 数据集上的不同初始重投影误差代价与初始空间误差代价比值 $r_c = C_S(\omega)/C_R$ 下的图像与激光数据融合精度（均方根误差）。
Table 4.3: Image and laser merging accuracies (RMSE) on NCT and FGT with different ratios of initial space error cost to initial reprojection error cost: $r_c = C_S(\omega)/C_R$.

$\lg(r_c)$	-3	-2	-1	0	1	2	3
NCT/mm	22.88	20.39	20.17	**19.42**	20.22	21.06	21.46
FGT/mm	33.02	32.29	28.24	**27.68**	30.76	35.62	39.98

式4.12中的优化问题在相机标定与激光对齐之间实现了较好的平衡，进而可以得到较好的图像与激光数据融合的结果。为验证上述猜想，本节分别定义初始重投影误差代价与初始空间误差代价如下：

$$\begin{aligned} C_R &= \sum_j \sum_k \rho\big(E_R(j,k)\big) \\ C_S(\omega) &= \omega \sum_i \Big(\sum_m \rho\big(E_S^G(i,m)\big) + \sum_n \rho\big(E_S^A(i,n)\big) \Big) \end{aligned} \tag{4.14}$$

初始误差代价表示该代价通过式4.12中待优化参数的初值计算得到。待优化参数的初值分别通过 SfM 以及激光数据粗略对齐过程中获取的，因此获取的初值均相对准确。本节将初始代价比 $C_S(\omega)/C_R$ 记为 r_c，不同的 r_c 在 NCT 与 FGT 数据集上的图像与激光数据融合精度如表4.3所示。注意，r_c 的值与 ω 的值成正比。

由表4.3可知，对于 NCT 与 FGT 数据集来说，随着 r_c 的增大，图像与激光数据融合的精度均为先增大后减小。当且仅当 r_c 的值处于一个较为合适的范围内时（$\lg(r_c) = -1, 0, 1$），图像与激光数据融合的精度较高，这验证了之前的猜想。因此，本章通过设置 ω 的值，使得初始空间误差与初始重投影误差相等：

$$\omega = \frac{\sum_j \sum_k \rho\big(E_R(j,k)\big)}{\sum_i \Big(\sum_m \rho\big(E_S^G(i,m)\big) + \sum_n \rho\big(E_S^A(i,n)\big) \Big)} \tag{4.15}$$

对比结果：最后，本节将本章的图像与激光数据融合方法与方法 Knapitsch et al. [13] 和 Schöps et al. [12] 进行了定量对比，对比结果如表4.4所示。其中 Coarse 为激光数据粗略对齐后的融合精度，而 $Fine$ 为经本章图像与激光数据融合后的融合精度。在文献 [13] 中，由图像生成的稠密点云通过一个带尺度的拓展 ICP 方法 [107] 对齐至激光点云。另外，由于方法 [43, 44] 与方法 [13] 原理上是类似的，因

表 4.4: 不同对比方法在 NCT 与 FGT 数据集上的图像与激光数据融合精度（均方根误差）。

Image and laser scan merging accuracy (RMSE) on NCT and FGT with different comparative methods.

对比方法	基线：Coarse	Knapitsch et al.	Schöps et al.	本章方法：Fine
NCT/mm	22.78	20.79	19.88	**19.42**
FGT/mm	32.96	30.47	30.64	**27.68**

此它们的融合精度不会高于 [13]。在文献 [12] 中，基于粗略对齐，首先通过点到面 ICP[108] 对用于激光数据对齐的相似变换进行优化，然后固定激光数据，通过改进的深度图像对齐方法 [109] 优化相机位姿。

由表4.4可知，在所有的对比方法中，本章方法（Fine）在基线（Coarse）的基础上提升最大。这是因为：（1）对于 Knapitsch et al. [13] 来说，由于经图像生成的稠密点云与激光点云在密度与噪声程度上的差异过大，两者之间很难通过基于 ICP 的方法实现精确对齐；（2）对于 Schöps et al. [12]，其融合精度很大程度上取决于进行激光数据对齐的 ICP 结果。然而本章中的激光数据仅作为图像的辅助，相邻扫描位置间的激光数据重叠区域十分有限。因此，在 NCT 与 FGT 数据集上，通过 ICP 不会得到十分精确的激光数据对齐结果以用于大幅度提升图像与激光数据融合的精度。

4.7 本章小结

本章提出了一个新颖的建筑场景重建流程，该流程利用了两种互相补足的元数据，图像与激光数据，以实现在数据采集效率以及重建精度与完整度上的较好平衡。本章方法以图像为主体用于完整覆盖场景，以激光数据为辅助用于针对弱纹理，弱光照以及结构复杂的区域。本章方法主要包括三步：图像采集，激光扫描以及图像与激光融合，通过该方法可以获取精确完整的场景重建结果。最后，通过在中国古代建筑数据集上的实验结果验证了本章方法的有效性。

第 5 章　融合迷你飞行器与机器人数据的室内建模

前面三章主要面向的研究对象是室外大规模场景（中国古代建筑）的精确、完整三维建模，当对室内场景进行建模时，融合多源数据也可以在精度与完整度方面对重建结果进行有效提升。本章中的研究对象是室内办公场景。在该场景中，重复结构、重复纹理以及场景自遮挡、互遮挡的情况更为严重。针对上述问题，本章提出了一种融合迷你飞行器与机器人数据的室内建模方法。

5.1　引言

室内场景三维重建在许多现实应用中起到了重要作用，例如室内导航、服务机器人、BIM 等。现有的室内场景重建方法可大致分为三类：（1）基于 LiDAR 的方法，（2）基于 RGB-D 相机的方法，（3）基于图像的方法。

尽管基于 LiDAR 的方法与基于 RGB-D 相机的方法均有着较高的精度，在重建较大的室内场景时，上述两种方法均存在成本较高，拓展性较差等问题。对于基于 LiDAR 的方法 [110, 111]，由于扫描视角限制导致场景遮挡难以避免，在进行扫描时往往需要多视角的激光扫描与点云对齐。对于基于 RGB-D 相机的方法 [112, 113]，由于传感器有效工作距离受限，需要采集、处理大量的数据。因此，上述方法在进行大规模室内场景重建时，均存在高成本、低效率的不足。

相对于基于 LiDAR 的方法与基于 RGB-D 相机的方法，尽管基于图像的方法成本更低，灵活性更强，这类方法也存在一些不足，如由于复杂场景、重复结构、缺乏纹理等导致的不完整、不精确的重建结果。由文献 [13] 可知，即使目前最先进的 SfM 与 MVS 技术，在规模较大，结构较复杂的室内场景中的重建效果仍不能令人满意。另外，一些基于图像的方法利用一些先验假设来处理室内场景重建问题，例如曼哈顿世界假设 [14]。尽管这些方法在有些时候能够取得较好的结果，但是，在不符合先验假设的情况下这些方法往往会导致错误的重建结果。

由于室内场景的复杂性，对于基于图像的方法实现场景完整重建需考虑如下两个问题。第一个是图像采集过程，即如何采集图像以完整、高效地覆盖室内场景。第二个是场景重建算法，即如何在 SfM 与 MVS 过程中融合不同视角的图像以获取完整、精确的重建结果。针对上述两问题，本章提出了一种新颖的基于图像的室内场景采集与重建流程。该流程用到了迷你飞行器与地面机器人并包含四个

主要步骤（见图5.1）：（1）首先采用一个迷你飞行器在室内采集图像，然后由飞行器图像获取表征室内场景的三角形网格，并将其用于为地面机器人定位导航的地图；（2）在飞行器地图中进行平面检测，获取地平面并用于地面机器人路径规划。然后，基于飞行器地图合成若干机器人视角图像，用于地面机器人的定位；（3）地面机器人进入室内场景进行机器人视角图像的采集。在机器人边运动边采集图像的同时，通过匹配采集的图像与合成的机器人视角图像，实现机器人的定位；（4）当地面机器人完成图像采集后，通过在基于图像的建模流程中融合迷你飞行器图像与地面机器人图像，实现室内场景的完整与精确建模。

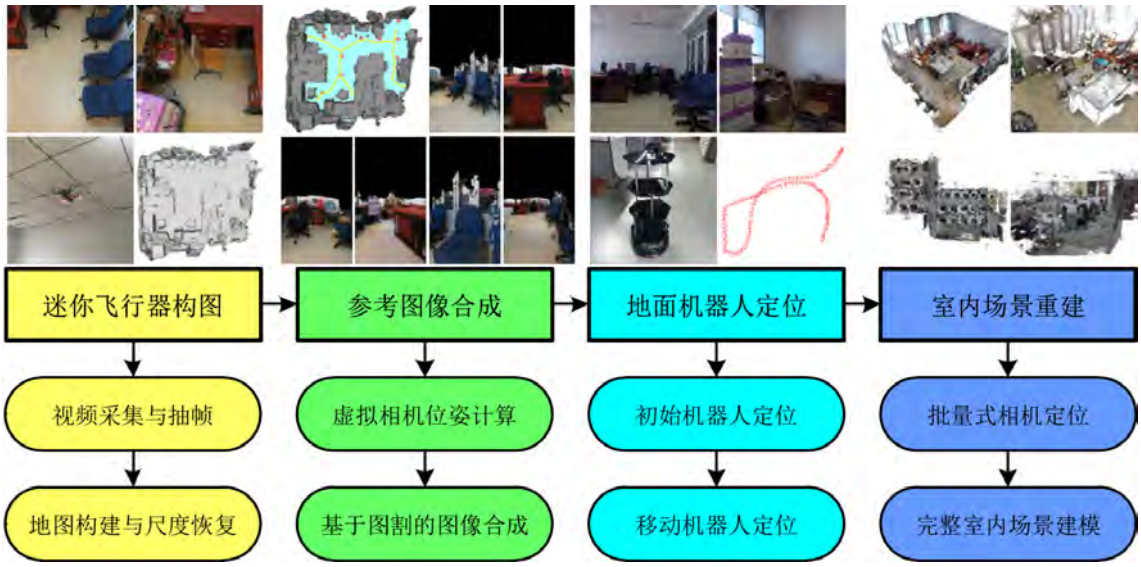

图 5.1: 本章方法流程图，主要包含四个步骤：（1）迷你飞行器构图；（2）参考图像合成；（3）地面机器人定位；（4）室内场景重建。
Figure 5.1: Pipeline of the proposed method in this chapter. It mainly contains four steps: (1) aerial map construction; (2) reference image synthesis; (3) ground robot localization; and (4) indoor scene reconstruction.

本章方法的主要贡献总结如下：

- 在整个系统流程中，只有飞行器图像采集过程中需要人工操作，后续的机器人图像采集以及室内场景建模过程均为全自动实现，这意味着本章方法的流程拓展性强，适用于大规模室内场景的采集与重建。
- 相比于地面机器人采集的图像，迷你飞行器采集的图像拥有更好的视角和更大的视场，这意味着相对于机器人图像，飞行器图像中的遮挡与误匹配问题会更小。因此，通过飞行器图像生成的地图能够更为可靠地用于后续的地面机器人定位过程中。

- 迷你飞行器拍摄的图像与地面机器人拍摄的图像相互补充并且能够完整覆盖室内场景。因此，通过融合飞行器与机器人图像，可以获取更为完整、精确的室内场景模型。

5.2 方法概述

本章中的室内场景重建流程图如图5.1所示，该流程主要包含四个步骤：（1）迷你飞行器构图；（2）参考图像合成；（3）地面机器人定位；（4）室内场景重建。各步骤具体细节在下文中详细介绍。

5.3 迷你飞行器构图

本章方法首先采用迷你飞行器在室内场景采集视频，并从视频中抽取一些图像。然后通过基于图像建模的流程对抽取的图像进行重建得到三维模型，并用将其作地面机器人定位的三维地图。

5.3.1 视频采集与抽帧

本章中，首先操控一架迷你飞行器在室内场景中采集自顶向下的视频，采集的视频分辨率为 1080p，帧率为 25FPS。由于迷你无人机尺寸小，灵活度高，十分适用于室内场景拍摄。举例说明，本章中采用的迷你飞行器为安装了稳定器与 4K 相机的 DJI Spark，其重量仅为 300g。另外，相对于机器人视角，从飞行器视角对室内场景进行拍摄不易受到场景遮挡的影响，因此采用迷你飞行器可以更加高效、完整覆盖场景。

给出采集的飞行器视频，可以通过 SLAM 系统，例如 ORB-SLAM[114]，构建飞行器地图。然而，本章采用离线的 SfM 技术 [73] 进行飞行器地图构建。这是因为：（1）在本章中飞行器地图用于地面机器人定位，因此没必要进行在线构建。（2）与容易产生场景漂移现象的 SLAM 相比，SfM 更加适用于大规模场景建模。可是，如果采用 SfM 进行飞行器地图构建时，显然不需要用到飞行器视频中的所有帧。因为飞行器视频帧中含有大量的冗余信息，这会严重降低 SfM 地图构建的效率。为解决上述问题，一个直接的办法就是在视频中每间隔固定的帧数抽取一帧，然后用抽取的视频帧进行地图构建。然而，这种做法仍存在一些缺点：（1）很难通过人工操作迷你飞行器在室内场景中实现稳定、恒速的视频采集，而这个问题在航线拐

角处会变得更加困难。(2) 由于室内场景中的纹理丰富程度是不一致的，因此对场景进行均匀覆盖也是不恰当的。为解决上述在飞行器地图构建过程中存在的问题，本章中采用了一种基于 BoW 模型的自适应视频抽帧方法，其过程详述如下。

BoW[115] 模型是一项在物体识别 [116]、基于图像的定位 [117]、SfM[118] 以及 SLAM[119] 等领域中广泛应用的图像检索技术。在 BoW 模型中，一幅图像可以表示为一个归一化向量 $\boldsymbol{v}_i$，而一对图像相似度可通过对应向量的点乘 $\boldsymbol{v}_i^T\boldsymbol{v}_i$ 表示。相邻图像之间过高的相似度会引入过多冗余信息，进而降低构图效率；而相邻图像之间过低的相似度则会导致图像之间连接性较差，构图不完整。因此，本章提出了一个从全体视频帧中自适应抽取子集的方法，在抽帧时该方法限定每个抽取的视频帧与其相邻的抽取的视频帧之间的相似度在一个合适的范围内。具体来说，本章方法首先通过 libvot 库 [120] 生成每一帧的归一化向量 $\boldsymbol{v}_i$，并将第一帧作为起始点。在抽帧过程中，假设当前第 i 帧已被抽取，本章方法获取该帧与其后续帧之间的相似度的得分：$\{s_{i,j}|j=i+1,i+2,\dots\}$，其中 $s_{i,j}=\boldsymbol{v}_i^T\boldsymbol{v}_i$，然后将 $s_{i,j}$ 与预设的相似度阈值 t 进行比较，本章中 $t=0.1$。假设 s_{i,j^*} 为 $\{s_{i,j}\}$ 中的第一个满足如下不等式：$s_{i,j}<t$，则第 j^*-1 帧（即第一个满足上述不等式的上一帧）为下一个抽取的视频帧。上述过程迭代进行，直至验证完所有视频帧。

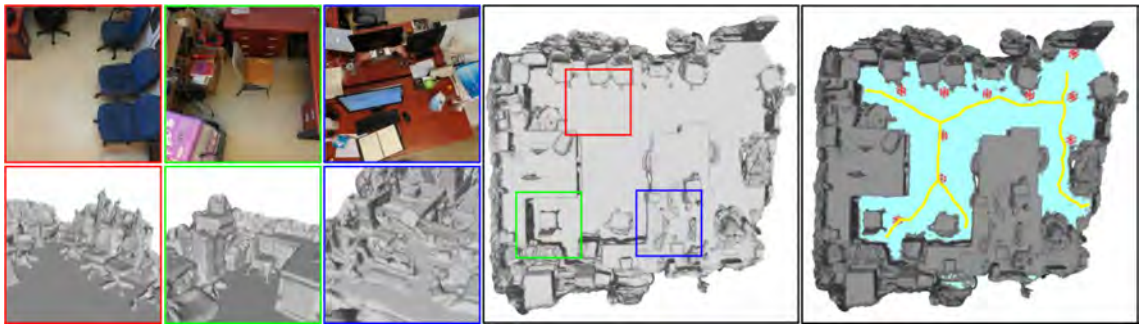

图 5.2: 图中前三列为示例飞行器图像及其对应的三维飞行器地图区域。第四列为整个三维飞行器地图。第五列为在飞行器地图上的机器人路径规划与虚拟相机位姿计算结果，其中地平面标为蓝色，规划路径标为黄色线段，虚拟相机位姿由红色棱锥表示。

Figure 5.2: The first three columns are the aerial image examples and their corresponding regions of the 3D aerial map. The forth column is the entire 3D aerial map. The fifth column is the robot path planning and virtual camera pose computation results on the aerial map, where the detected ground plane is shown blue, the planned path is denoted as yellow line segments, and the computed virtual camera poses are represented by red pyramids.

5.3.2 地图构建与尺度恢复

给出经上述方法抽取的飞行器视频帧，通过一套标准的基于图像建模流程构建飞行器地图，该流程包括：（1）SfM[9]，（2）MVS[75]，（3）表面重建 [7]。另外，由于室内接收不到 GPS 信号，本章方法通过 GCP 将飞行器地图缩放至其真实物理尺寸。图5.2为一个经 271 幅抽取的视频帧重建得到的飞行器地图。

5.4 参考图像合成

上文中构建的飞行器地图在后续过程中起到了两个作用：第一个是为地面机器人规划路径并在机器人移动过程中进行定位；第二个是在室内场景重建过程中有助于飞行器与机器人图像的融合。上述两个过程均需要建立机器人图像与飞行器地图之间的二维到三维的点的对应关系。为获取上述对应点，一个可能有效的解决方案是直接匹配飞行器与机器人图像。然而，由于这两种图像在视角上差异巨大，直接对其进行匹配是十分困难的。在此，本章方法通过由飞行器地图合成机器人视角参考图像的方式解决上述问题。参考图像经如下两步进行合成：虚拟相机位姿计算以及基于图割的图像合成。

5.4.1 虚拟相机位姿计算

用于参考图像合成的虚拟相机位姿基于室内场景的地平面进行计算，本章中飞行器地图的地平面通过一个基于 RANSAC 的形状检测方法 [121] 进行检测（见图5.2）。虚拟相机位姿分两步进行计算，先计算位置后计算朝向。

5.4.1.1 位置计算

本章方法求取地平面的二维包围盒并将其划分成方形栅格，栅格的大小决定了虚拟相机的数量。为在定位精度与效率上达到平衡，本章中将栅格边长设为 $1m$。对于每个栅格，当其中的地平面面积占栅格总面积的比例大于 50% 时，本章方法认为该栅格为放置虚拟相机的有效栅格。虚拟相机位置设为有效栅格的中心并有着高度为 h 的高程偏移量（见图5.2）。h 的值由机器人相机的高度决定，在本章中其值设为 $1m$。

5.4.1.2 朝向计算

在得到虚拟相机位置以后，为实现对场景的全方向观测，需要在每个虚拟相机位置放置多个光心相同、朝向不同的虚拟相机 [12]。本章中，由于安装在地面机器人上的相机的光轴近似平行于地平面，在此只生成水平朝向的虚拟相机。另外，为消除机器人与合成图像之间的透视投影失真，本章方法将虚拟相机的视场（内参数）设为与机器人相机接近。在本章中，每个虚拟相机位置上放置 6 个虚拟相机，虚拟相机之间的偏航角夹角为 60°。

另外，用于地面机器人运动的路径也要通过检查的地平面进行规划。路径规划问题在机器人领域研究广泛 [122]，且已有许多有效的解决方案 [123, 124]。由于本章方法并非聚焦于规划地面机器人的最优路径，在此将检测的地平面的骨架用作机器人路径，该骨架通过中轴变换法 [125] 进行提取（见图5.2）。

5.4.2 基于图割的图像合成

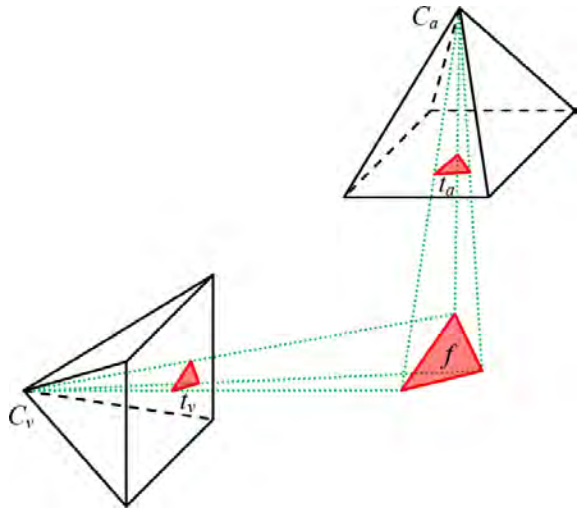

图 5.3: 基于网格的图像合成示意图。其中 f 为一个三维空间面片，其在飞行器 C_a 与虚拟 C_v 相机上的二维投影三角形分别记作 t_a 与 t_v。图像合成的原理是将 t_a 经过 f 变至 t_v。
Figure 5.3: Schematic diagram of mesh based image synthesis. f is a spatial facet, whose projective 2D triangles on an aerial camera (C_a) and a virtual camera (C_v) are denoted as t_a and t_v respectively. The image is synthesized by warping t_a to t_v through f.

与3.4.2节类似，此处的图像合成也是借助于空间连续的网格（见图5.3）。具体来说，本章方法先获取每个飞行器与虚拟相机的可见网格。然后，对于每个虚拟相机，将其可见网格投影至该相机上形成二维三角形集合。在进行虚拟图像合成时，对于虚拟图像中的一个特定的二维三角形来说，本章方法需要基于如下三个因素

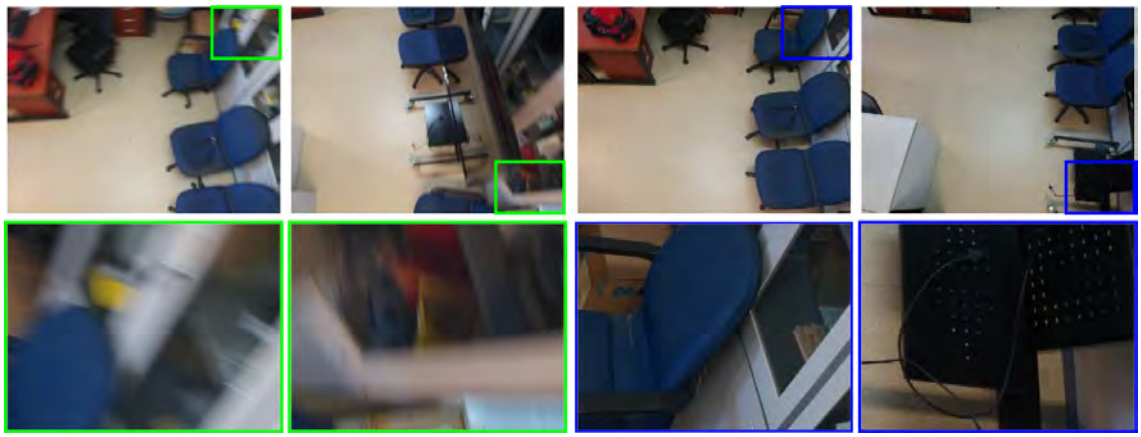

图 5.4: 局部特征尺度与图像清晰度之间的关系。左边两列：两幅局部特征尺度中值最大的图像。右边两列：两幅局部特征尺度中值最小的图像。第二行为第一行（绿色/蓝色）矩形区域的放大图像。

Figure 5.4: Correlation between feature scale and image sharpness. Left two columns: Images with two largest median feature scales. Right two columns: Images with two smallest median feature scales. The second row is the enlarged images of the rectangular regions in the first row.

确定采用哪一幅飞行器图像进行变换以填充此区域：（1）可见性，对于此二维三角形对应的三维空间面片，选取的飞行器图像应有较好的视角与较近的视距；（2）清晰度，由于从室内飞行器视频抽帧得到的图像中有一部分清晰度较差，要在其中选取足够清晰的飞行器图像；（3）一致性，虚拟图像中相邻的三角形应尽可能由同一幅飞行器图像进行合成以保持合成图像的一致性。本章中，可见性因素通过空间面片在飞行器图像上的投影面积衡量（越大越好），而清晰度因素通过飞行器图像局部特征尺度的中值衡量（越小越好，具体见图5.4）。基于上述描述，本章中的图像合成问题可归结为多标签优化问题，定义如下：

$$E(l) = E_{data}(l) + E_{smooth}(l) = \sum_{t_i \in \mathcal{T}} D_i(l_i) + \sum_{\{t_i,t_j\} \in \mathcal{N}} V_{i,j}(l_i, l_j) \tag{5.1}$$

其中，$\mathcal{T}$ 为虚拟相机可见的三维空间网格投影得到的二维三角形集合，t_i 为其中的第 i 个三角形；$\mathcal{N}$ 为投影三角形的公共边集合；l_i 为 t_i 的标签，即飞行器图像序号。当对应 t_i 的空间面片在第 l_i 个飞行器图像中可见时，数据项 $D_i(l_i) = \sigma_{l_i}/A_{l_i}$，其中 σ_{l_i} 为第 l_i 个飞行器图像中局部特征的尺度中值而 A_{l_i} 为对应 t_i 的空间面片在第 l_i 个飞行器图像中的投影面积；否则的话 $D_i(l_i) = \alpha$，其中 α 为一个较大的常量（本章中 $\alpha = 10^4$）以惩罚这种情况。当 $l_i = l_j$ 时，平滑项 $V_{i,j}(l_i, l_j) = 0$；否则 $V_{i,j}(l_i, l_j) = 1$。定义于式5.1的优化问题可通过图割算法 [126–128] 进行高效求解。

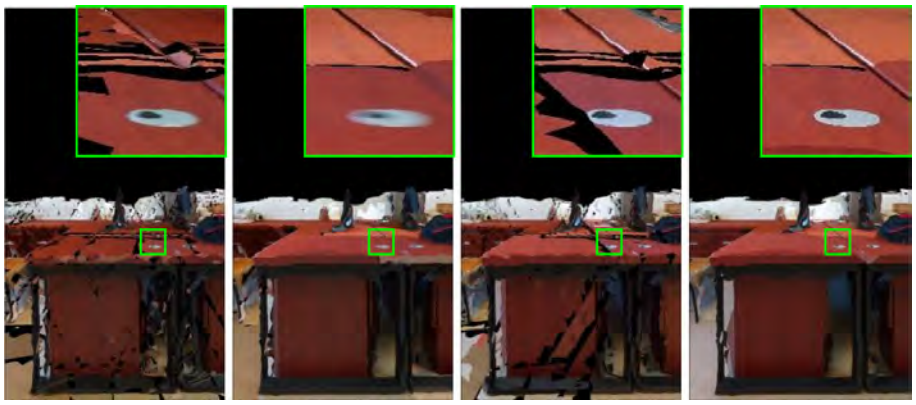

图 5.5: 不同配置下基于图割的图像合成结果。从左到右：既不考虑清晰度因素，又不考虑一致性因素；只考虑一致性因素；只考虑清晰度因素；既考虑清晰度因素，又考虑一致性因素的图像合成结果。每幅图右上角的大矩形为图中小矩形的方大图。

Figure 5.5: One example of graph cut based image synthesis. From left to right: Image synthesis result by considering neither the sharpness nor the consistency factor; only the consistency factor; only the sharpness factor; and both the sharpness and the consistency factors. The larger rectangle at the upper-right corner of each image is the enlarged version of the smaller one in each image.

图 5.6: 另外的一些图像合成结果以及类似视角下的机器人图像。

Figure 5.6: Some other synthetic image examples and their corresponding ground images with similar viewpoints.

为阐明清晰度因素与一致性因素的影响，此处在四种不同配置下在其中一个虚拟相机上进行了图像合成，结果如图5.5所示。由5.5图可知，清晰度因素使得合成图像更为清楚而一致性因素使得合成图像中孔洞及锐边更少。另外，图5.6给出了另外的一些图像合成结果以及类似视角下的机器人图像。尽管仍有些难以避免的合成错误情况，合成图像与其对应的机器人图像在公共可见区域有着较大的相似性，这验证了本章中图像合成方法的有效性。本节中的合成图像将用作机器人机器人定位的参考数据库。

5.5 地面机器人定位

在将地面机器人放置于室内场景中时，机器人将沿着规划路径运动并自动采集机器人视角视频。如果机器人仅通过其内置传感器，例如轮子编码器与 IMU，进行定位的话，它将不会严格按照规划的路径运动。这是因为机器人内置传感器存在累积误差的问题，这种问题对于安装在消费级机器人上的低成本传感器来说尤为明显。因此，机器人的位姿需要通过视觉定位的方式进行修正，而在本章中通过匹配合成与机器人图像实现视觉定位。

5.5.1 初始机器人定位

通过对机器人相机采集视频的第一帧进行定位，可以获取机器人在飞行器地图中的初始位置，并将该位置作为机器人后续运动的起点。上述初始定位可通过匹配第一帧图像与所有合成图像或者通过语义树 [120] 检索得到的 k 个最相似的合成图像实现。本章中使用的是基于图像检索的方法，且 $k = 30$。需要注意的是，尽管在此合成了机器人视角图像，机器人图像与合成图像在光照、视角等方面仍有较大区别，常用的 SIFT 特征不足以应对。本章采用的为 ASIFT 特征。

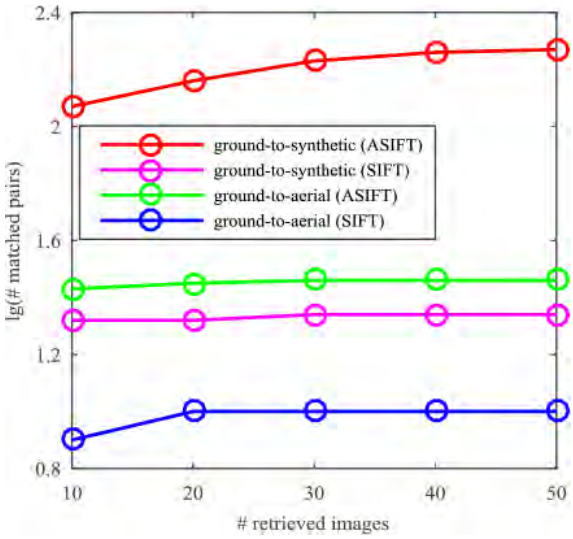

图 5.7: 图像匹配结果。其中，x 轴为检索图像数量，y 轴为匹配图像对数量的对数。
Figure 5.7: Image matching results. The x-axis is the number of retrieved images and the y-axis is the logarithm of the number of matched image pairs.

为验证本章图像合成方法的有效性并对 SIFT 特征与 ASIFT 特征的性能进行比较，分别采用 SIFT 特征与 ASIFT 特征进行了合成与机器人图像匹配以及飞行器与机器人图像匹配。其中，机器人图像也是通过5.3.1节介绍的抽帧方法从机器

人机器人采集的视频中抽取获得。在进行图像匹配时，检索了不同数量的与当前机器人图像最近似的合成图像与飞行器图像。当经过基本矩阵验证后的匹配点数仍大于 16 时，这对图像是匹配的。图像匹配结果如图5.7所示。由图5.7可知，采用 ASIFT 进行合成与机器人图像匹配得到的匹配对数分别大约是采用 ASIFT 进行飞行器与机器人图像匹配，采用 SIFT 进行合成与机器人图像匹配以及采用 SIFT 进行飞行器与机器人图像匹配的 6 倍，8 倍与 19 倍。

给出第一帧机器人图像与检索的合成图像之间的二维匹配点，可以通过光线投射的方式在飞行器地图上获取对应的三维空间点。这样一来可以采用基于 PnP 的方法实现第一帧机器人图像的定位。具体来说，给定二维到三维对应点与机器人相机内参数，相机位姿通过 RANSAC 采用不同的 PnP 算法进行求解。采用的 PnP 算法包括 P3P[129]，AP3P[130] 与 EPnP[131]。当上述算法对应的内点数有至少一种超过 16 个时，此次位姿估计为一次成功的估计，并将该相机的位姿定为 PnP 结果中内点数量最多的那一个。在本章 RANSAC 过程中，一共进行了 500 次随机抽样，且将距离阈值设为 $4px$。

5.5.2 移动机器人定位

地面机器人在室内场景中运动并采集视频时，可以通过轮子里程计对其粗略定位。本章通过匹配机器人与合成图像将地面机器人全局式地定位至飞行器地图上以修正机器人粗略定位结果。此处只对抽取的机器人视频帧，而非全部视频帧进行位姿修正。这是因为：（1）地面机器人在室内运动相对缓慢，在较短时间内不会严重偏离规划路径。（2）每次进行全局视觉定位需要耗时大约 $0.5s$，且时间主要耗在 ASIFT 特征提取上。需要注意的是，对于某些抽取的视频帧，由于用于 PnP 的内点数量不足，视觉定位并不能一直成功。

假设上一次成功定位的机器人图像的位置与朝向分别记为 $\boldsymbol{c}_A$ 与 $\boldsymbol{n}_A$，而当前待定位的机器人图像通过轮子里程计得到的粗略位置与朝向分别记为 $\boldsymbol{c}_B$ 与 $\boldsymbol{n}_B$。在此基于粗略定位结果，而非基于图像检索的方法查找当前机器人图像的候选匹配合成图像。该方法的示意图如图所示5.8。当合成图像满足如下两个条件时，本章方法将对其与当前机器人图像进行匹配：（1）合成图像位于圆心为 $\boldsymbol{c}_B$，半径为 r_B 的圆中，其中 $r_B = \max(\|\boldsymbol{c}_B - \boldsymbol{c}_A\|, \beta)$ 且 $\beta = 2m$；（2）合成图像朝向与 $\boldsymbol{n}_B$ 的夹角小于 90°。在此用到一个可变半径 r_B 的原因是随着机器人的运动，通过机器人内置传感器获取的相对位姿的漂移会越来越严重。在对当前机器人图像与得

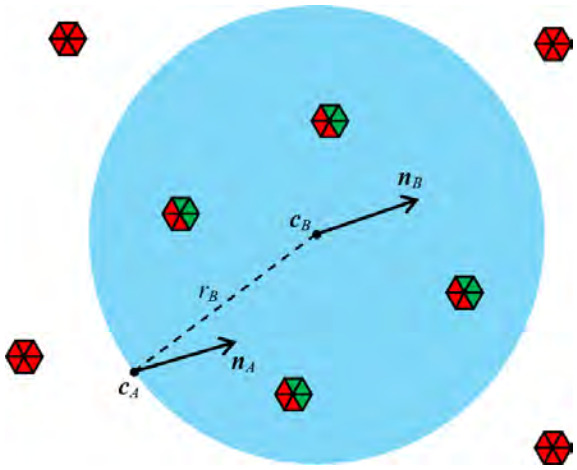

图 5.8: 机器人运动过程中候选匹配合成图像查找示意图。$\boldsymbol{c}_A$ 与 $\boldsymbol{n}_A$ 为上一次成功定位的机器人图像的位置与朝向。$\boldsymbol{c}_B$ 与 $\boldsymbol{n}_B$ 为当前的机器人图像的粗略位置与朝向。图中蓝色的圆表示查找范围，该圆圆心为 $\boldsymbol{c}_B$，半径为 r_B。图中三角形表示虚拟相机位姿，绿色的三角形表示选中的合成图像而红色的三角形表示未选中的合成图像。

Figure 5.8: Schematic diagram of finding potential matched synthetic images during robot movement. $\boldsymbol{c}_A$ and $\boldsymbol{n}_A$ are the location and orientation of the last successfully localized extracted ground frame. $\boldsymbol{c}_B$ and $\boldsymbol{n}_B$ are the coarse location and orientation of the currently extracted ground frame. The blue circle denotes the searching region (with center $\boldsymbol{c}_B$ and radius r_B). The triangles denote the virtual camera poses. The green triangles are the selected synthetic images while the red ones are not.

到的候选匹配合成图像进行匹配之后，当前机器人图像采用类似5.5.1节中的方法，通过基于 PnP 的 RANSAC 的方法实现定位。如果定位结果在位置和朝向上与粗略定位结果偏差足够小（本章中位置偏差小于 $5m$，朝向偏差小于 $30°$），当前机器人图像定位成功。此时认为机器人的位姿已通过当前定位成功的机器人图像全局修正，并将轮子里程计中的位姿重置为当前基于视觉的定位结果。注意，未定位成功的机器人图像将在后续室内场景重建过程中重新定位。

5.6 室内场景重建

在机器人定位与视频采集后，并非所有从机器人视频抽取的帧均已成功定位至飞行器地图。然而，为获取完整的室内场景重建结果，需要定位并融合所有由（飞行器与机器人）视频中抽取得到的图像。在此，本章方法首先提出了一种批量式定位之前未成功定位的机器人图像的流程。然后，本章方法将机器人与合成图像匹配内点连入原始特征点轨迹中，并通过 BA 实现飞行器与机器人点云的融合。最后，本章方法通过融合飞行器与机器人图像以获取完整、稠密的室内场景重建结果。

5.6.1 批量式相机定位

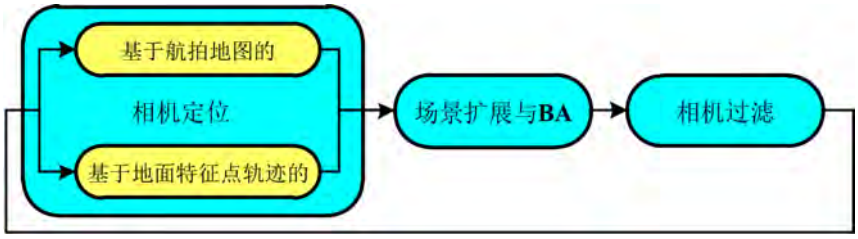

图 5.9: 批量式相机定位流程图，该流程以循环的形式进行，每个循环中包括三个步骤：（1）相机定位；（2）场景扩展与 BA；（3）相机过滤。
Figure 5.9: Flow chart of the batched camera localization. It goes in loops and each loop contains three steps: (1) camera registration; (2) scene expansion and BA; and (3) camera filtering.

为定位之前未成功定位的机器人图像，在此本章提出了一种批量式相机定位流程。与文献 [132] 类似，在每个相机定位循环中，本章方法尽量定位更多的相机。然而，此处用于相机定位的二维到三维对应点中的三维空间点并不像文献 [132] 中仅包括在 SfM 过程中重建得到的空间点，还包括通过光线投射与飞行器地图（三维网格）相交得到的空间点。每个批量式相机定位循环中包括三个步骤：（1）相机定位，（2）场景扩展与 BA，（3）相机过滤，其流程图如图5.9所示。在进行批量式相机定位之前，本章方法先对从机器人视频中抽帧得到的图像进行匹配并将匹配点连接成特征点轨迹 [133]。对于至少有两幅已成功定位的可见图像的特征点轨迹，本章方法通过三角测量的方式 [9] 求取其空间坐标。

5.6.1.1 相机定位

在此有两种方式获取二维三维对应点以定位当前未定位成功的机器人图像：（1）飞行器地图，对于当前未成功定位的机器人图像中的二维特征点，可以获取其在成功定位的图像中的匹配点。然后从成功定位的相机光心向这些匹配点投射射线，投射的射线与飞行器地图的交点即为当前未成功定位的机器人图像中的二维特征点对应的三维空间点。（2）机器人特征点轨迹，给出当前通过三角测量得到的机器人特征点轨迹，可以通过先前机器人图像之间的匹配结果获取当前未成功定位的机器人图像中对应的二维特征点。当前未定位成功的机器人相机可利用上述两种二维三维对应点通过基于 PnP 的 RANSAC 的方法实现定位，而定位结果采用两结果中内点数多的那一个。在此，将通过两种二维三维对应点实现相机定位的方法与只用其中任意一种的方法进行了比较，结果如图5.10所示。由图5.10可知，经

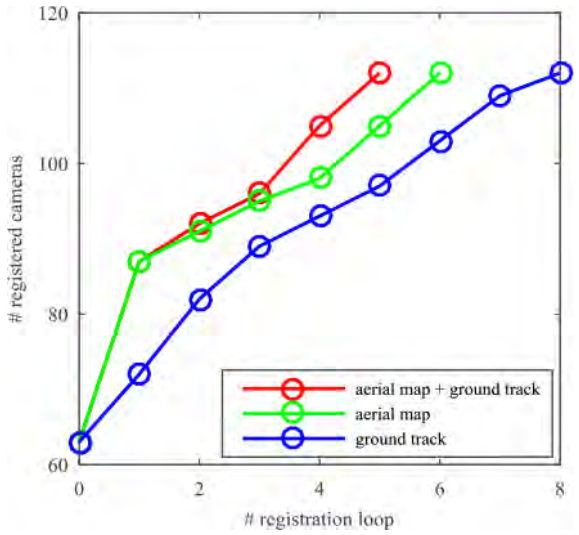

图 5.10: （1）基于飞行器地图与机器人特征点轨迹，（2）仅基于飞行器地图，（3）仅基于机器人特征点轨迹的批量式相机定位结果。图中 x 轴为批量式相机定位循环次数，y 轴为成功定位的相机数量。注意，当 $x = 0$ 时对应的 y 值为在5.5.2节中成功定位的相机数量。
Figure 5.10: Batched camera registration results of using (1) both aerial map and ground tracks, (2) aerial map alone, and (3) ground tracks alone. The x-axis is the times of camera registration loops and the y-axis is the number of registered cameras. Note that the value of y when $x = 0$ is the number of registered cameras during robot localization in Sec. 5.5.2.

过若干迭代循环，三种方法均可定位同样数量的相机。然而，本章中通过两种二维三维对应点实现相机定位的方法所需的迭代循环次数最少（仅需 5 次，而其他两种方法分别需要 6 次与 8 次）。

5.6.1.2 场景扩展与 BA

在相机定位之后，本章方法根据新定位的相机对机器人特征点轨迹进行三角测量以实现场景扩展。为提高相机位姿与场景点的精度，在三角测量后对已定位的机器人相机地位姿与三角测量得到的机器人特征点轨迹的空间位置通过 BA 进行优化。

5.6.1.3 相机过滤

考虑到方法的鲁棒性，本章方法在 BA 后对定位成功的相机加入了一步相机过滤的操作。若在本次迭代循环中新定位成功的相机，经 BA 优化后的位置或朝向与其粗略定位结果（轮子里程计获取的定位结果）偏差较大（位置偏差大于 $5m$ 或朝向偏差大于 30°）的话，此定位结果不可靠并将其滤除。注意，在当前迭代循

环中滤除的相机在后续迭代循环中仍可成功定位。

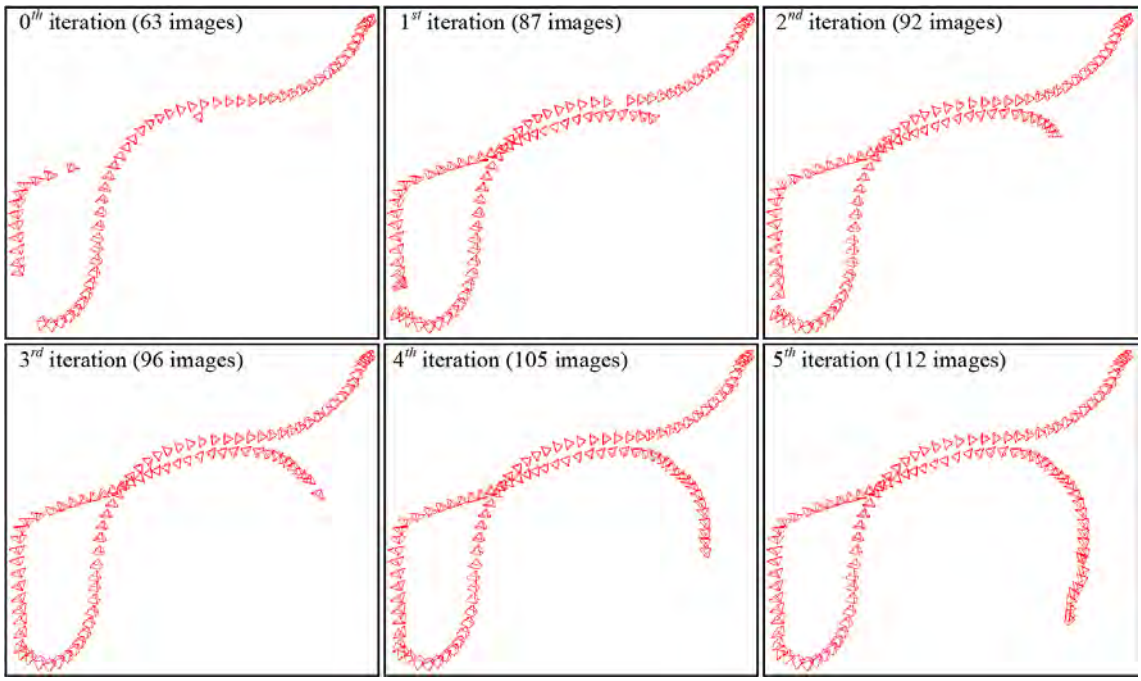

图 5.11: 批量式相机定位过程，其中红色棱锥表示定位成功的相机位姿。第 0 次迭代表示在5.5.2节中的相机定位结果。
Figure 5.11: Batched camera localization process. The red pyramids denote the localized camera poses. The 0-th iteration is the robot visual localization result in Sec. 5.5.2.

上述三个步骤迭代进行，直至所有相机均成功定位或者不再有相机可以成功定位。批量式相机定位的过程如图5.11所示。

5.6.2 完整室内场景建模

在对机器人视频抽帧得到的图像批量式定位之后，本章方法利用所有飞行器与机器人图像重建获取完整的室内场景。首先，将机器人与合成图像匹配点连入原始的飞行器与机器人特征点轨迹中以生成跨视图的约束。然后，通过 BA 对飞行器与机器人点云进行融合。最后，通过融合飞行器与机器人图像获取室内场景的完整、稠密模型。

5.6.2.1 飞行器与机器人特征点轨迹生成

为通过 BA 融合飞行器与机器人点云，需要引入飞行器与机器人图像之间的约束。在此，上述跨视图约束可通过由5.5节中获取的机器人与合成图像匹配点生成的飞行器与机器人特征点轨迹提供。匹配的机器人图像特征点可通过查询其序

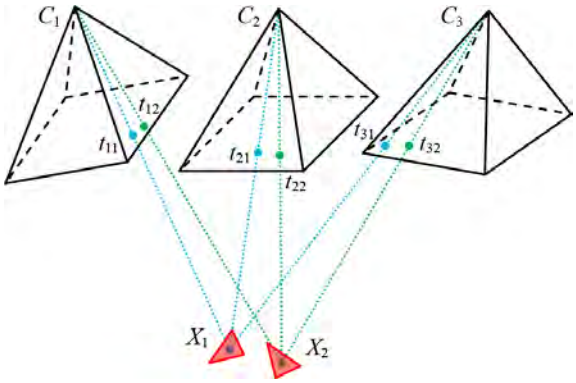

图 5.12: 针对飞行器视图的飞行器与机器人特征点轨迹生成示意图。其中，$C_i(i=1,2,3)$ 为飞行器相机，$X_j(j=1,2)$ 为对应于匹配的合成图像特征点的空间点，t_{ij} 为点 X_j 在相机 C_i 上的投影，$t_{1j}-t_{2j}-t_{3j}(j=1,2)$ 为第 j 个跨飞行器视图的特征点轨迹。
Figure 5.12: Schematic diagram of ground-to-aerial tracks generation for aerial views. $C_i(i=1,2,3)$ are the aerial cameras. $X_j(j=1,2)$ are the spatial points corresponding to the matched synthetic feature points. t_{ij} is the projection of X_j in C_i. $t_{1j}-t_{2j}-t_{3j}(j=1,2)$ is the j-th track across the aerial views.

号较为便捷地连入原始机器人特征点轨迹中。但是，尽管合成图像由飞行器图像生成，想要将匹配的合成图像特征点连入原始飞行器特征点轨迹中却没那么容易。这是因为用于与机器人图像匹配的合成图像特征点是在合成图像上重新提取得到的。本章通过光线投射与点投影的方式将机器人与合成图像匹配点拓展至飞行器视图，该过程的示意图如图5.12所示。具体来说，先通过光线投射的方式在飞行器地图上获取匹配的合成图像特征点对应的空间点，然后将获取的空间点投影至其可见飞行器图像上以生产飞行器与机器人特征点轨迹。

5.6.2.2 图像融合与模型重建

接下来，本章方法通过 BA 对连接生成的飞行器与机器人特征点轨迹，原始的（飞行器与机器人）特征点轨迹，所有（飞行器与机器人）相机的内外参数进行全局优化。然后，采用方法 [75] 利用飞行器与机器人图像进行稠密重建以获取室内场景的稠密模型。由于在优化过程中引入了跨飞行器与机器人视图的约束，且稠密重建过程中融合了飞行器与机器人图像，重建得到的模型比仅用单一来源的图像重建得到的模型更加完整、精确。

5.7 实验结果

本节对本章中提出的室内场景采集与重建流程进行了评测。首先，介绍了用于采集飞行器与机器人元数据的实验设备，以及采集到的两组室内场景数据集。然后，在这两组数据集上对本章方法进行了测试。

5.7.1 数据集

图 5.13: 本章实验中用到的元数据采集设备。从左到右：机器人上的 TurtleBot，空中的 DJI Spark，桌面上的 DJI Spark。
Figure 5.13: Data acquisition equipments in the experiments of this chapter. From left to right: the TurtleBot on the ground, the DJI Spark in the air, and the DJI Spark on the desk.

表 5.1: Room 与 Hall 数据集元数据。
Table 5.1: Meta-data of Room and Hall datasets.

数据集	Room	Hall
飞行器视频长度 $/s$	218	494
机器人视频长度 $/s$	61	113
覆盖面积 $/m^2$	30	130

由于目前几乎没有针对室内场景的飞行器与机器人图像公开数据集，在此，通过自建室内场景数据集进行方法评测。具体来说，采用 DJI Spark 进行飞行器视角场景采集，采用安装在 TurtleBot 上的 GoPro HERO4 进行机器人视角场景采集，元数据采集设备如图5.13所示。采集的飞行器与机器人元数据的形式均为分辨率为 1080p，帧率为 25FPS 的视频。采集的两个室内场景数据集分别叫做 Room 与 Hall。一些关于 Room 与 Hall 数据集的信息如表5.1所示。Room 与 Hall 数据集中的示例飞行器图像与生成的三维飞行器地图分别如图5.2与图5.14所示。由图5.2与图5.14可知，Hall 数据集的飞行器地图相对于 Room 数据集的飞行器地图质量更

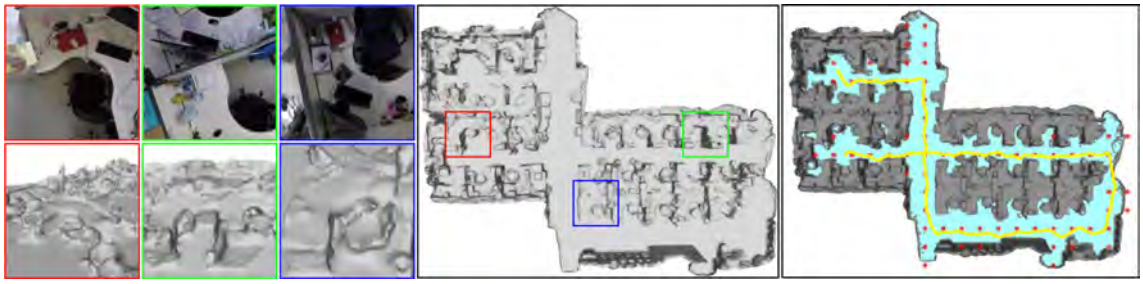

图 5.14: Hall 数据集中的示例飞行器图像与生成的三维飞行器地图。图中前三列为示例飞行器图像及其对应的三维飞行器地图区域。第四列为整个三维飞行器地图。第五列为在飞行器地图上的机器人路径规划与虚拟相机位姿计算结果，其中地平面标为蓝色，规划路径标为黄色线段，虚拟相机位姿由红色棱锥表示。

Figure 5.14: The first three columns are the aerial image examples and their corresponding regions of the 3D aerial map of the Hall dataset. The forth column is the entire 3D aerial map. The fifth column is the robot path planning and virtual camera pose computation results on the aerial map, where the detected ground plane is shown blue, the planned path is denoted as yellow line segments, and the computed virtual camera poses are represented by red pyramids.

差且规模更大。然而，由后续的方法测评章节可知，本章方法在上述两个数据集上均可取得预期的结果，这说明本章方法有着较好的鲁棒性与可拓展性。

另外，在 Room 与 Hall 数据集上的虚拟相机位姿计算与机器人路径规划结果分别展示于图5.2与图5.14的最右侧。如图所示，通过本章方法，用于虚拟相机位姿计算与机器人路径规划的地平面可以成功检测，且生成的虚拟相机与机器人路径较为均匀地覆盖了室内场景。本章的虚拟相机位姿计算方法在 Room 与 Hall 数据集上分别生成了 60 与 384 个虚拟相机。

5.7.2 自适应抽帧结果

通过本章的自适应抽帧方法，分别从 Room 数据集的飞行器与机器人视频中抽取了 271 与 112 帧图像，从 Hall 数据集的飞行器与机器人视频中抽取了 721 与 250 帧图像。为验证本章中抽帧方法的有效性，在 Hall 数据集的飞行器视频上对本章方法与等间隔抽帧方法进行了对比实验。采用本章的自适应抽帧方法在长度为 494s，帧率为 25FPS 的视频上抽取得到了 721 帧图像，对于等间隔抽帧方法，每隔 17 帧抽取 1 帧图像（$494 \times 25/721 \approx 17$），共计抽取 730 帧图像。然后，本节将两种不同抽帧方法得到的视频帧通过开源 SfM 系统 COLMAP[9] 进行相机标定，结果如图5.15所示。由图5.15可知，由于相比等间隔的方法，本章方法抽取的视频帧连接性更好，因此通过对其进行重建，可获得一致的飞行器地图。另外，图5.15中

图 5.15: Hall 数据集飞行器视频上的本章抽帧方法与等间隔抽帧方法对比实验结果。左图：自适应抽取的视频帧的 COLMAP 结果，其中 98.61%($\frac{711}{721}$) 的视频帧成功标定。中图和右图：等间隔抽取的视频帧的 COLMAP 结果，其中 97.12%($\frac{367+342}{730}$) 的视频帧成功标定，但断开为两部分。中图与右图分别对应着左图中的绿色与蓝色矩形区域。左图与右图中的黑色圆展示了在同一拐角处的对比结果。

Figure 5.15: Comparative results between the proposed adaptive frame extraction method in this chapter and the method of frame extraction with constant interval on the aerial frames of the Hall dataset. Left: COLMAP result on the adaptively extracted aerial frames, where 98.61%($\frac{711}{721}$) frames are successfully registered. Middle and right: COLMAP result on the constantly extracted aerial frames, where 97.12%($\frac{367+342}{730}$) frames are registered in two separated models. The middle and right figures correspond to the green and blue rectangles in the left figure respectively. The black circles in the left and right figures show the comparative results of the same turning corner.

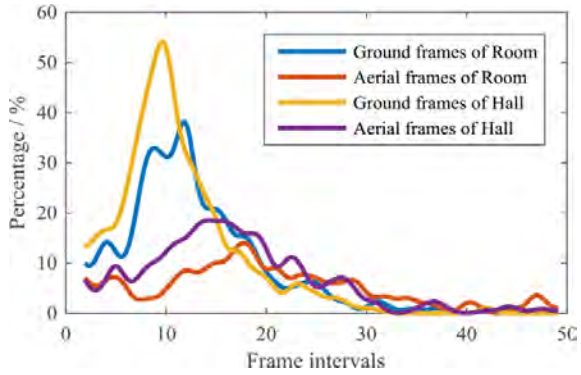

图 5.16: 本章抽帧方法在 Room 与 Hall 的飞行器与机器人视频上的抽取帧的间隔分布。

Figure 5.16: Interval distributions of the proposed adaptive frame extraction method in this chapter on ground/aerial frames of the Room and Hall datasets.

的黑圆表明，为获取更加完整的飞行器地图，需要在拐角处对视频进行更加密集的抽帧操作。最后，通过本章抽帧方法在 Room 与 Hall 的飞行器与机器人视频上的抽取帧的间隔分布如图所示5.16。由图5.16可知，自适应抽取视频帧的间隔近似服从泊松分布（不同视频，分布的期望值不同）。

5.7.3 机器人相机定位结果

为验证本章的批量式相机定位以及飞行器与机器人图像融合方法，在此对批量式相机定位（5.6.1节）与飞行器与机器人图像融合（5.6.2节）后的相机定位结果以及 COLMAP 结果进行了定性与定量比较。需要注意的是，对于 COLMAP 来说，机器人相机位姿并未进行初始化，仅通过图像本身进行标定，即在5.5.2节中借助飞行器地图的相机定位结果并未提供给 COLMAP 用作先验。

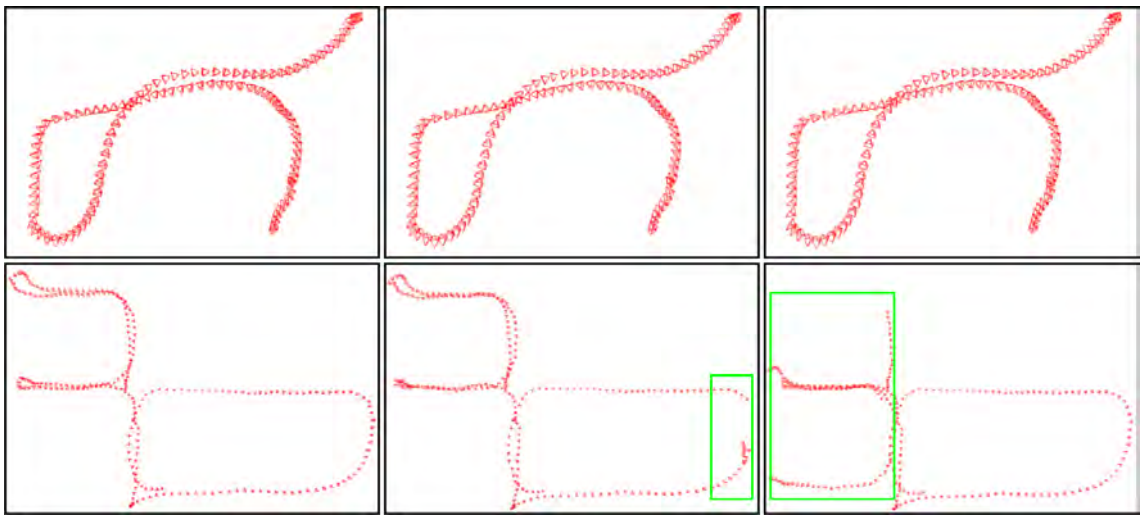

图 5.17: 机器人相机定位的定性对比结果。第一行：Room 数据集结果；第二行：Hall 数据集结果。从左到右：飞行器与机器人图像融合后的结果；机器人相机批量式定位后的结果；COLMAP 标定结果。图中绿色矩形标示出了错误的相机位姿。
Figure 5.17: Qualitative comparison results of ground camera localization. First row: results of the Room dataset; second row: results of the Hall dataset; From left to right: results after image merging; results after batched camera localization; and calibration results of COLMAP. The green rectangles denote the incorrect camera poses.

定性对比结果如图5.17所示，由图5.17可知，对于 Room 数据集，通过三种对比方法获取的相机位姿较为类似，这是由于 Room 数据集的场景结构较为简单。而对于 Hall 数据集来说，通过 COLMAP 计算得到的相机轨迹在场景的左边部分有着明显的错误。这是由于重复纹理与弱纹理导致机器人图像之间的匹配结果包含较多的匹配外点，这样的话会导致增量式 SfM 系统产生较为明显的场景漂移现象。相比之下，对于批量式相机定位来说，由于部分机器人图像已初步定位至飞行器地图，其定位结果仅存在一些较为轻微的场景漂移情况。并且，上述错误的相机姿态均在后续的飞行器与机器人图像融合阶段修正过来。这是由于，在图像融合时的全局优化中引入了连接生成的飞行器与机器人特征点轨迹。上述结果表明，通过融合飞行器与机器人图像对机器人相机进行定位相比于仅用机器人图像来说更为鲁棒。

为了对机器人相机定位结果进行定量测评，需要提供相机位姿（位置与朝向）的真值。本章提出了一种给出机器人相机位姿近似真值的方法。该测评方法基于如下假设：在室内场景中，飞行器图像的相机定位与场景重建精度高于机器人图像的。上述假设基于的事实是相对于机器人图像，飞行器图像有着更好的视角以及更少的遮挡。该假设将在后续章节进行验证。对于本章的测评方法，具体来说，对每个数据集，随机选取 10 幅机器人图像，对于每幅选出的图像，人工获取若干其与飞行器图像的二维匹配点。然后，可以通过光线投射的方式获取选取的机器人图像与飞行器地图之间的二维三维对应点。最后，通过 PnP 算出选取的机器人图像相对于飞行器地图的位置、朝向，并将其用作近似的真值。

表 5.2: 机器人相机定位定量对比结果。表中结果为真值与定位结果在相机位置与朝向上的 RMSE。
Table 5.2: Quantitative comparison results of ground camera localization. The results in the table are the RMSE between the ground truths and the localization results in location and orientation.

数据集	Room			Hall		
方法	批量定位	图像融合	COLMAP	批量定位	图像融合	COLMAP
位置 RMSE/m	0.0177	**0.0124**	0.0235	0.2529	**0.1403**	0.9979
朝向 RMSE/°	0.3591	**0.1995**	0.8871	1.0334	**0.6394**	2.0167

基于相机位姿真值，本节对机器人相机定位结果进行了定量的测评，结果如表5.2所示。从表5.2中可以看到与图5.17中的定性对比结果类似的现象，即在 Room 数据集上各对比方法取得的精度较为接近，而在 Hall 数据集上定位精度差异较大（图像融合 > 批量定位 >COLMAP）。另外，相对于位置误差，三个方法在朝向误差上的差异相对较小，上述现象表明在机器人相机定位的过程中，朝向估计比位置估计更为鲁棒。

5.7.4 室内场景重建结果

最后，本节对本章中的室内场景重建算法进行了定性与定量测评。本节比较了本章中的室内重建结果与仅采用飞行器或机器人图像进行重建的结果，定性比较结果如图5.18所示。需要注意的是：（1）对于本章中的室内重建算法，采用的相机位姿为融合飞行器与机器人图像之后的相机位姿；（2）对于仅采用机器人图像的方法，采用的相机位姿为经过批量式相机定位之后的相机位姿；（3）对于仅采用飞行器相机的方法，采用的相机位姿为经过 SfM 估计得到的相机位姿。由图5.18可

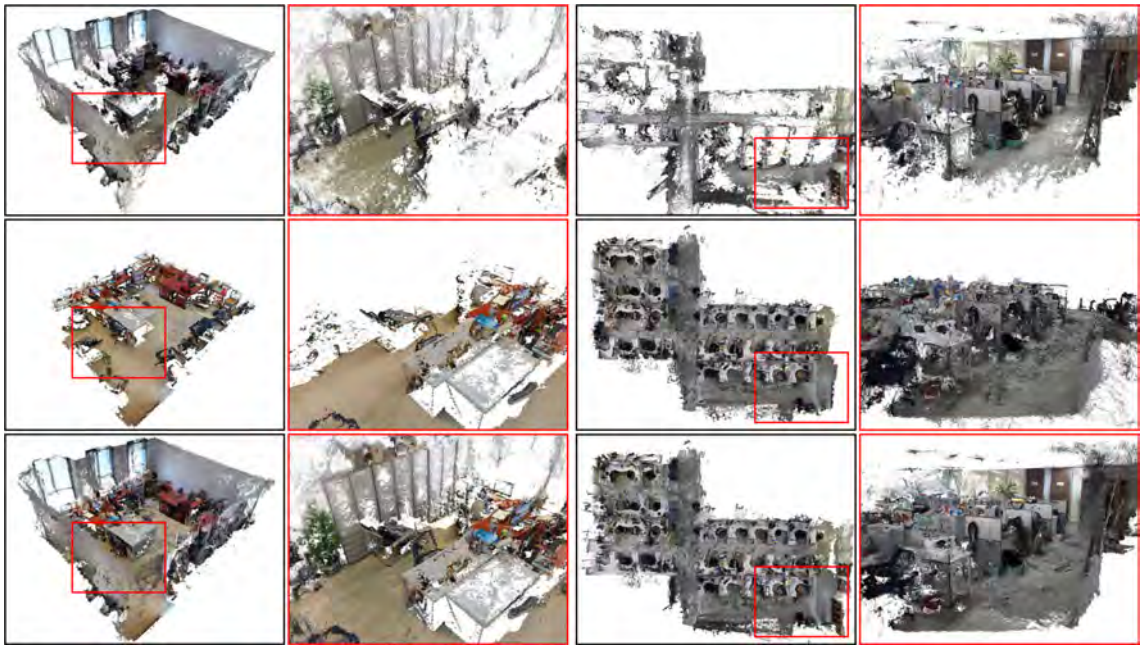

图 5.18: 室内场景重建定性结果。第一列：Room 数据集结果；第二列：第一列中红色矩形区域的放大图；第三列：Hall 数据集结果；第四列：第三列中红色矩形区域的放大图。从上到下：仅用机器人图像，仅用飞行器图像，利用融合的飞行器与机器人图像的结果。
Figure 5.18: Qualitative indoor scene reconstruction results. First column: results of the Room dataset; second column: enlarged versions of red rectangles in the first column; third column: results of the Hall dataset; fourth column: enlarged versions of red rectangles in the third column. From top to bottom: Results of using ground images alone, aerial images alone, and merged ground and aerial images.

知，尽管由于遮挡与弱纹理情况的存在，重建结果中仍不可避免地缺失了部分区域，相对于仅采用单独一种图像进行重建，通过融合飞行器与机器人图像的室内重建结果更为完整。

图 5.19: Room 数据集上的用于定量评价室内场景重建结果的实验设置。其中，红色、绿色与蓝色线段分别为机器人到机器人、飞行器到飞行器与飞行器到机器人空间线段示例。
Figure 5.19: Experimental setup for the quantitative evaluation of the indoor scene reconstruction results on the Room dataset. The red, green and blue line segments denote the ground-to-ground, aerial-to-aerial and ground-to-aerial segment examples respectively.

表 5.3: 室内场景重建定量结果。细节见正文。
Table 5.3: Quantitative results of indoor scene reconstruction. See text for more details.

数据集	Room			Hall		
模型	飞行器模型	机器人模型	融合模型	飞行器模型	机器人模型	融合模型
飞行器到飞行器 RMSE/m	0.0357	–	0.0203	0.0563	–	0.0324
机器人到机器人 RMSE/m	–	0.0811	0.0430	–	0.3121	0.1104
飞行器到机器人 RMSE/m	–	–	0.0341	–	–	0.0578

另外，本节通过如下方式对重建结果进行了定量评测。对于 Room 与 Hall 数据集，分别在重建模型中获取了 30 条空间线段。这 30 条空间线段分为三组：（1）前 10 条线段的两个端点仅在飞行器相机中可见；（2）中间 10 条线段的两个个端点仅在机器人相机中可见；（3）后 10 条线段的两个端点，一个仅在飞行器图像中可见，另一个仅在机器人图像中可见。因此，可以从飞行器模型与融合模型上获取第一组线段的长度，可以从机器人模型与融合模型上获取第二组线段的长度，而仅可以从融合模型上获取第三组线段的长度。通过比较第一（二）组线段在飞行器（机器人）模型与在融合模型上的长度差异，可以对飞行器（机器人）模型在图像融合前后的融合精度进行评测。第三组线段用于表示飞行器与机器人图像融合精度。在此采用 Leica X310 激光测距仪对上述空间线段的真值进行测量，该测距仪测量范围为 $0.05 \sim 120m$，测量精度为 $\pm 1mm$。对于每组线段，本节在模型上获取其长度并与真值比较以求得 RMSE，如表5.3所示。由表5.3可知：（1）Room 数据集上的重建精度高于 Hall 数据集；（2）在室内场景中，飞行器模型重建精度高于机器人模型；（3）通过融合飞行器与机器人图像，室内场景的重建精度提高了。因此，通过本章中的融合飞行器与机器人的方法，可以获取更加精确、完整的室内场景模型。

5.8 本章小结

本章提出了一个新颖的基于图像的室内场景重建流程，用于重建室内场景的图像包括采用迷你飞行器从空中采集的以及采用机器人从地面采集的。在本章流程首先构建了一个三维飞行器地图用引导机器人在室内场景中行进并采集机器人视角图像。然后，本章方法对飞行器与机器人图像进行融合，并通过融合后的图像生成完整、精确的室内场景模型。本章的室内场景重建流程兼顾采集效率与重建精度，并且，本章方法在两组室内场景数据集上验证了该流程的有效性及鲁棒性。

第 6 章　总结与展望

6.1　工作总结

本文针对传统基于图像的大规模场景三维重建方法难以兼顾重建精度与完整度的问题，提出了一系列融合多源数据的场景重建算法，实现了室内外大规模场景的精确、完整重建。本文的主要工作总结如下：

1. 基于稠密点云的航拍与地面点云对齐方法

采用由粗到精的流程实现航拍与地面稠密点云的对齐。为提高点云对齐的精度与效率，通过对地面稠密点云进行投影的方式实现航拍视角图像的合成。并且在点云对齐的过程中从图像选取、合成与匹配三方面进行了改进，使得合成的图像分布均匀，噪声较小，可得到更多的匹配内点。实验结果表明本章方法可有效地实现航拍与地面模型的对齐，且相比于其他方法，本方法在对齐精度与效率方面均表现更好。

2. 基于稀疏点云的航拍与地面点云融合方法

针对基于稠密点云投影的点云对齐方法效率较低，合成图像噪声大、有孔洞等问题，采用基于稀疏网格诱导单应的方式合成航拍视角图像；针对基于图像的建模中难以避免的场景漂移问题，采用捆绑调整的方式实现航拍与地面点云融合；针对航拍图像与合成图像匹配外点较多的问题，采用基于几何一致性检验和几何模型验证的方式对匹配外点进行过滤。实验结果表明，提出的航拍与地面图像匹配方法在召回率、精度与效率方面优于其它对比方法；提出的航拍与地面点云融合方法在精度与效率方面优于其它对比方法。

3. 融合图像与激光数据的精确完整建模方法

采用融合图像与激光数据的方式，实现大规模场景的精确、完整建模。首先对场景进行图像采集获取稀疏点云及网格，根据得到的稀疏网格，自动规划激光扫描站点的位置。在求取扫描站点位置时，综合考虑场景结构复杂程度、纹理丰富程度以及站点分布情况，通过贪婪算法近似求解。通过投影激光点云合成航拍与地面视角图像，并将其与采集图像进行匹配，采用广义捆绑调整的方式实现图像与激光数据的融合。实验结果表明，该方法能通过融合图像与激光数据有效地实现大规模场景的精确、完整建模。

4. **融合迷你飞行器与机器人数据的室内建模方法**

采用迷你飞行器采集图像构图，用于地面机器人路径规划并辅助机器人定位。采用基于图割的方式合成虚拟的机器人视角图像用于与匹配机器人图像实现机器人的全局定位。通过融合迷你飞行器与地面机器人图像的方式实现室内场景的精确、完整建模。实验结果表明，该方法可实现室内场景中地面机器人的精确定位以及场景的完整建模。

6.2 工作展望

尽管本文在融合多源数据进行大规模场景三维重建方面开展了一系列工作并取得了部分结果。为进一步提升重建效果，应从如下四个方面继续开展相关工作：

1. **多源数据评测数据集**

目前用于评测融合多源数据进行大规模场景三维重建的公开数据集十分紧缺；另外，用于定量评测融合多源数据的重建结果的评测标准也同样缺乏。这就导致本文中的实验大部分采用了自建数据集以及自定义的近似定量评价指标。为促进相关领域的发展，包含多源数据，以精确、完整重建大规模场景为目标并且拥有相对应地真值与评测指标的数据集亟待提出。

2. **激光为主，图像为辅的室内场景三维重建**

本文提出的融合图像与激光数据的建模方法已图像为主，激光为辅进行场景三维重建，该方法通常可以取得较好的结果。然而，当场景结构过于复杂，光照、纹理过弱时，基于图像的重建方法难以获取较好的重建结果，上述情况在室内场景中更为显著。此时，可采用激光为主，图像为辅的重建流程。首先采用激光进行场景扫描，基于扫描结果获取重建缺失区域并对该区域进行图像采集。最后通过融合图像与激光数据的方式获取场景精确。完整的重建结果。

3. **融合 RGB 图像、RGB-D 图像、LiDAR 数据的大规模场景三维重建**

尽管通过融合 RGB 图像与 LiDAR 数据可以获得相对更加精确、完整的大规模场景重建结果，然而，由于存在弱纹理区域与复杂场景导致的遮挡情况，仅对上述两种类型数据进行融合得到的重建结果在完整度方面仍有进一步的提升空间。本文拟通过分析融合 RGB 图像与 LiDAR 数据得到的重建结果以自动获取需进一步提升完整度的区域。之后，本文拟采用 Kinect 采集该区域的 RGB-D 图像，并将其与先前重建结果进行融合以获取更加精确、完整的重建结果。

4. **融合结构信息与语义信息的室内机器人定位与场景重建**

仅通过图像局部特征在室内场景中进行机器人定位与场景重建存在较大的局限性。当前存在一些利用高层结构信息 [52, 55] 或者语义信息 [51, 134] 进行视觉定位的方法。该类高层信息虽然精度较低，但鲁棒性更强。后续工作拟将上述高层信息融合至本文的室内机器人定位与场景重建框架之中，以提升系统的鲁棒性与有效性。

参考文献

[1] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91–110.

[2] MOREL J M, YU G. ASIFT: a new framework for fully affine invariant image comparison[J]. SIAM Journal on Imaging Sciences, 2009, 2(2): 438–469.

[3] RUSU R B, BLODOW N, BEETZ M. Fast point feature histograms (FPFH) for 3D registration[C]//IEEE International Conference on Robotics and Automation (ICRA). 2009: 3212–3217.

[4] SHAN Q, WU C, CURLESS B, et al. Accurate geo-registration by ground-to-aerial image matching[C]//International Conference on 3D Vision (3DV). 2014: 525–532.

[5] SNAVELY N, SEITZ S M, SZELISKI R. Modeling the world from internet photo collections[J]. International Journal of Computer Vision, 2008, 80(2): 189–210.

[6] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(8): 1362–1376.

[7] VU H H, LABATUT P, PONS J P, et al. High accuracy and visibility-consistent dense multiview stereo[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(5): 889–901.

[8] CUI Z, TAN P. Global structure-from-motion by similarity averaging[C]//IEEE International Conference on Computer Vision (ICCV). 2015: 864–872.

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

[10] CHATTERJEE A, GOVINDU V M. Robust relative rotation averaging[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 958–972.

[11] STRECHA C, VON HANSEN W, GOOL L V, et al. On benchmarking camera calibration and multi-view stereo for high resolution imagery[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008: 1–8.

[12] SCHÖPS T, SCHÖNBERGER J L, GALLIANI S, et al. A multi-view stereo benchmark with high-resolution images and multi-camera videos[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2538–2547.

[13] KNAPITSCH A, PARK J, ZHOU Q Y, et al. Tanks and temples: Benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 78:1–78:13.

[14] FURUKAWA Y, CURLESS B, SEITZ S M, et al. Manhattan-world stereo[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009: 1422–1429.

[15] CUI Z, GU J, SHI B, et al. Polarimetric multi-view stereo[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 369–378.

[16] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346–359.

[17] ARANDJELOVIĆ R, ZISSERMAN A. Three things everyone should know to improve object retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012: 2911–2918.

[18] ZITNICK C L. Binary coherent edge descriptors[C]//European Conference on Computer Vision (ECCV). 2010: 170–182.

[19] KUSHNIR M, SHIMSHONI I. Epipolar geometry estimation for urban scenes with repetitive structures[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(12): 2381–2395.

[20] WU C, FRAHM J M, POLLEFEYS M. Detecting large repetitive structures with salient boundaries[C]//European Conference on Computer Vision (ECCV). 2010: 142–155.

[21] BANSAL M, SAWHNEY H S, CHENG H, et al. Geo-localization of street views with aerial image databases[C]//ACM International Conference on Multimedia (MM). 2011: 1125–1128.

[22] BANSAL M, DANIILIDIS K, SAWHNEY H. Ultra-wide baseline facade matching for geo-localization[C]//European Conference on Computer Vision Workshops (ECCVW). 2012: 175–186.

[23] WOLFF M, COLLINS R T, LIU Y. Regularity-driven building facade matching between aerial and street views[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1591–1600.

[24] ZHOU L, ZHU S, SHEN T, et al. Progressive large scale-invariant image matching in scale space[C]//IEEE International Conference on Computer Vision (ICCV). 2017: 2381–2390.

[25] LI X, LARSON M, HANJALIC A. Pairwise geometric matching for large-scale

object retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 5153–5161.

[26] FISCHLER M A, BOLLES R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381–395.

[27] JOHNSON A E, HEBERT M. Using spin images for efficient object recognition in cluttered 3D scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(5): 433–449.

[28] GUO Y, SOHEL F, BENNAMOUN M, et al. Rotational projection statistics for 3D local surface description and object recognition[J]. International Journal of Computer Vision, 2013, 105(1): 63–86.

[29] GUO Y, BENNAMOUN M, SOHEL F, et al. A comprehensive performance evaluation of 3d local feature descriptors[J]. International Journal of Computer Vision, 2016, 116(1): 66–89.

[30] BESL P J, MCKAY N D. A method for registration of 3-D shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239–256.

[31] WU C, CLIPP B, LI X, et al. 3D model matching with viewpoint-invariant patches (VIP)[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2008: 1–8.

[32] KAMINSKY R S, SNAVELY N, SEITZ S M, et al. Alignment of 3d point clouds to overhead images[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW). 2009: 63–70.

[33] BÓDIS-SZOMORÚ A, RIEMENSCHNEIDER H, GOOL L V. Efficient volumetric fusion of airborne and street-side data for urban reconstruction[C]//International Conference on Pattern Recognition (ICPR). 2016: 3204–3209.

[34] ZHOU Y, SHEN S, GAO X, et al. Accurate mesh-based alignment for ground and aerial multi-view stereo models[C]//IEEE International Conference on Image Processing (ICIP). 2017: 2627–2631.

[35] UMEYAMA S. Least-squares estimation of transformation parameters between two point patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(4): 376–380.

[36] LIU L, STAMOS I. A systematic approach for 2D-image to 3D-range registration in urban environments[C]//IEEE International Conference on Computer Vision (ICCV). 2007: 1–8.

[37] BILA Z, REZNICEK J, PAVELKA K. Range and panoramic image fusion into a textured range image for culture heritage documentation[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2013, II-5/W1: 31–36.

[38] SIRMACEK B, LINDENBERGH R C, MENENTI M. Automatic registration of Iphone images to laser point clouds of the urban structures using shape features [J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2013, II-5/W2: 265–270.

[39] STAMOS I, ALIEN P K. Automatic registration of 2-D with 3-D imagery in urban environments[C]//IEEE International Conference on Computer Vision (ICCV). 2001: 731–736.

[40] LI Y, ZHENG Q, SHARF A, et al. 2D-3D fusion for layer decomposition of urban facades[C]//IEEE International Conference on Computer Vision (ICCV). 2011: 882–889.

[41] NEX F, GERKE M, REMONDINO F, et al. ISPRS benchmark for multi-platform photogrammetry[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, II-3/W4: 135–142.

[42] BASTONERO P, DONADIO E, CHIABRANDO F, et al. Fusion of 3D models derived from TLS and image-based techniques for CH enhanced documentation[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2014, II-5: 73–80.

[43] RUSSO M, MANFERDINI A M. Integration of image and range-based techniques for surveying complex architectures[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2014, II-5: 305–312.

[44] ALTUNTAS C. Integration of point clouds originated from laser scaner and photogrammetric images for visualization of complex details of historical buildings[J]. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, XL-5/W4: 431–435.

[45] FRUEH C, ZAKHOR A. 3D model generation for cities using aerial photographs and ground level laser scans[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2001: II–31–II–38.

[46] THRUN S. Probabilistic robotics[M]. [S.l.]: MIT Press, 2005.

[47] MAJDIK A L, ALBERS-SCHOENBERG Y, SCARAMUZZA D. MAV urban lo-

calization from google street view data[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2013: 3979–3986.

[48] LI A, MORARIU V I, DAVIS L S. Planar structure matching under projective uncertainty for geolocation[C]//European Conference on Computer Vision (ECCV). 2014: 265–280.

[49] VISWANATHAN A, PIRES B R, HUBER D. Vision based robot localization by ground to satellite matching in gps-denied situations[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2014: 192–198.

[50] MATEI B C, VALK N V, ZHU Z, et al. Image to LIDAR matching for geotagging in urban environments[C]//IEEE Workshop on Applications of Computer Vision (WACV). 2013: 413–420.

[51] OZCANLI O C, DONG Y, MUNDY J L. Geo-localization using volumetric representations of overhead imagery[J]. International Journal of Computer Vision, 2016, 116(3): 226–246.

[52] WANG X, VOZAR S, OLSON E. FLAG: Feature-based localization between air and ground[C]//IEEE International Conference on Robotics and Automation (ICRA). 2017: 3178–3184.

[53] VANDAPEL N, DONAMUKKALA R R, HEBERT M. Unmanned ground vehicle navigation using aerial ladar data[J]. International Journal of Robotics Research, 2006, 25(1): 31–51.

[54] FORSTER C, PIZZOLI M, SCARAMUZZA D. Air-ground localization and map augmentation using monocular dense reconstruction[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2013: 3971–3978.

[55] SURMANN H, BERNINGER N, WORST R. 3D mapping for multi hybrid robot cooperation[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017: 626–633.

[56] MUJA M, LOWE D G. Scalable nearest neighbor algorithms for high dimensional data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36 (11): 2227–2240.

[57] SCARAMUZZA D. 1-point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints[J]. International Journal of Computer Vision, 2011, 95(1): 74–85.

[58] LIU Z, MARLET R. Virtual line descriptor and semi-local graph matching method

for reliable feature correspondence[C]//British Machine Vision Conference (BMVC). 2012: 16.1–16.11.

[59] LIU M Y, TUZEL O, VEERARAGHAVAN A, et al. Fast directional chamfer matching[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 1696–1703.

[60] CRISPELL D, MUNDY J, TAUBIN G. A variable-resolution probabilistic three-dimensional model for change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(2): 489–500.

[61] POMERLEAU F, MAGNENAT S, COLAS F, et al. Tracking a depth camera: Parameter exploration for fast ICP[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2011: 3824–3829.

[62] XIAO J, ADLER B, ZHANG J, et al. Planar segment based three-dimensional point cloud registration in outdoor environments[J]. Journal of Field Robotics, 2013, 30 (4): 552–582.

[63] HARTLEY R I, ZISSERMAN A. Multiple view geometry in computer vision[M]. Second ed. [S.l.]: Cambridge university press, 2004.

[64] HARTLEY R I. In defense of the eight-point algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(6): 580–593.

[65] NISTER D. An efficient solution to the five-point relative pose problem[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(6): 756–770.

[66] LEVENBERG K. A method for the solution of certain non-linear problems in least squares[J]. Quarterly of applied mathematics, 1944, 2(2): 164–168.

[67] MARQUARDT D W. An algorithm for least-squares estimation of nonlinear parameters[J]. Journal of the society for Industrial and Applied Mathematics, 1963, 11(2): 431–441.

[68] ZACH C. Robust bundle adjustment revisited[C]//European Conference on Computer Vision (ECCV). 2014: 772–787.

[69] RAGURAM R, FRAHM J, POLLEFEYS M. Exploiting uncertainty in random sample consensus[C]//IEEE International Conference on Computer Vision (ICCV). 2009: 2074–2081.

[70] CHUM O, MATAS J, KITTLER J. Locally optimized RANSAC[C]//Joint Pattern Recognition Symposium. 2003: 236–243.

[71] RAGURAM R, CHUM O, POLLEFEYS M, et al. USAC: A universal framework for

random sample consensus[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 2022–2038.

[72] CUI H, SHEN S, 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.

[73] CUI H, GAO X, SHEN S, et al. HSfM: Hybrid structure-from-motion[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2393–2402.

[74] HÄNE C, ZACH C, COHEN A, et al. Dense semantic 3D reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1730–1743.

[75] SHEN S. Accurate multiple view 3D reconstruction using patch-based stereo for large-scale scenes[J]. IEEE Transactions on Image Processing, 2013, 22(5): 1901–1914.

[76] MIKOLAJCZYK K, SCHMID C. A performance evaluation of local descriptors [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615–1630.

[77] ZHENG E, DUNN E, JOJIC V, et al. Patchmatch based joint view selection and depthmap estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014: 1510–1517.

[78] BARTELSEN J, MAYER H, HIRSCHMÜLLER H, et al. Orientation and dense reconstruction of unordered terrestrial and aerial wide baseline image sets[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2012, I-3: 25–30.

[79] MANCINI F, DUBBINI M, GATTELLI M, et al. Using unmanned aerial vehicles (UAV) for high-resolution reconstruction of topography: The structure from motion approach on coastal environments[J]. Remote Sensing, 2013, 5(12): 6880–6898.

[80] ROTTENSTEINER F, SOHN G, GERKE M, et al. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 93(7): 256–271.

[81] AGARWAL S, FURUKAWA Y, SNAVELY N, et al. Building Rome in a day[J]. Communications of the ACM, 2011, 54(10): 105–112.

[82] HEINLY J, SCHONBERGER J L, DUNN E, et al. Reconstructing the world* in six days[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3287–3295.

[83] GARLAND M, HECKBERT P S. Surface simplification using quadric error metrics [C]//ACM SIGGRAPH. 1997: 209–216.

[84] GREENE N, KASS M, MILLER G. Hierarchical z-buffer visibility[C]//ACM SIGGRAP. 1993: 231–238.

[85] RAY H, PFISTER H, SILVER D, et al. Ray casting architectures for volume visualization[J]. IEEE Transactions on Visualization and Computer Graphics, 1999, 5 (3): 210–223.

[86] JÉGOU H, DOUZE M, SCHMID C. Improving bag-of-features for large scale image search[J]. International Journal of Computer Vision, 2010, 87(3): 316–336.

[87] ZEISL B, SATTLER T, POLLEFEYS M. Camera pose voting for large-scale image-based localization[C]//IEEE International Conference on Computer Vision (ICCV). 2015: 2704–2712.

[88] PHILBIN J, CHUM O, ISARD M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007: 1–8.

[89] SCHÖNBERGER J L, RADENOVIĆ F, CHUM O, et al. From single image query to detailed 3D reconstruction[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 5126–5134.

[90] LIN W Y, LIU S, JIANG N, et al. RepMatch: Robust feature matching and pose for reconstructing modern cities[C]//European Conference on Computer Vision (ECCV). 2016: 562–579.

[91] SHEN S, HU Z. How to select good neighboring images in depth-map merging based 3D modeling[J]. IEEE Transactions on Image Processing, 2014, 23(1): 308–318.

[92] UMMENHOFER B, BROX T. Global, dense multiscale reconstruction for a billion points[J]. International Journal of Computer Vision, 2017, 125(1): 82–94.

[93] PARK J, SINHA S N, MATSUSHITA Y, et al. Robust multiview photometric stereo using planar mesh parameterization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1591–1604.

[94] ZHENG Q, SHARF A, WAN G, et al. Non-local scan consolidation for 3D urban scenes[C]//ACM SIGGRAPH. 2010: 94:1–94:9.

[95] NAN L, SHARF A, ZHANG H, et al. Smartboxes for interactive urban reconstruction[C]//ACM SIGGRAPH. 2010: 93:1–93:10.

[96] VANEGAS C A, ALIAGA D G, BENES B. Automatic extraction of manhattan-

world building masses from 3D laser range scans[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(10): 1627–1637.

[97] LI M, WONKA P, NAN L. Manhattan-world urban reconstruction from point clouds [C]//European Conference on Computer Vision (ECCV). 2016: 54–69.

[98] COHEN A, SCHÖNBERGER J L, SPECIALE P, et al. Indoor-outdoor 3D reconstruction alignment[C]//European Conference on Computer Vision (ECCV). 2016: 285–300.

[99] SOUDARISSANANE S, LINDENBERGH R. Optimizing terrestrial laser scanning measurement set-up[J]. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2011, XXXVIII-5/W12: 127–132.

[100] WUJANZ D, NEITZEL F. Model based viewpoint planning for terrestrial laser scanning from an economic perspective[J]. ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016, XLI-B5: 607–614.

[101] JIA F, LICHTI D. A comparison of simulated annealing, genetic algorithm and particle swarm optimization in optimal first-order design of indoor TLS networks [J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2017, IV-2/W4: 75–82.

[102] JIA F, LICHTI D. An efficient, hierarchical viewpoint planning strategy for terrestrial laser scanner networks[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, IV-2: 137–144.

[103] DESERNO M. How to generate equidistributed points on the surface of a sphere [J]. P.-If Polymerforshung (Ed.), 2004: 99.

[104] PLESS R. Using many cameras as one[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2003: II–587–93.

[105] LEE G H, FAUNDORFER F, POLLEFEYS M. Motion estimation for self-driving cars with a generalized camera[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 2746–2753.

[106] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679–698.

[107] LIN B, TAMAKI T, RAYTCHEV B, et al. Scale ratio ICP for 3D point clouds with different scales[C]//IEEE International Conference on Image Processing (ICIP). 2013: 2217–2221.

[108] CHEN Y, MEDIONI G. Object modelling by registration of multiple range images [J]. Image and Vision Computing, 1992, 10(3): 145–155.

[109] ZHOU Q Y, KOLTUN V. Color map optimization for 3D reconstruction with consumer depth cameras[J]. ACM Transactions on Graphics, 2014, 33(4): 155:1–155:10.

[110] MURA C, MATTAUSCH O, VILLANUEVA A J, et al. Automatic room detection and reconstruction in cluttered indoor environments with complex room layouts[J]. Computers & Graphics, 2014, 44(7): 20–32.

[111] PREVITALI M, BARAZZETTI L, BRUMANA R, et al. Towards automatic indoor reconstruction of cluttered building rooms from point clouds[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2014, II-5: 281–288.

[112] CHOI S, ZHOU Q Y, KOLTUN V. Robust reconstruction of indoor scenes[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 5556–5565.

[113] DAI A, NIE¦ NER M, ZOLLHÖFER M, et al. Bundlefusion: Real-time globally consistent 3D reconstruction using on-the-fly surface reintegration[J]. ACM Transactions on Graphics, 2017, 36(3).

[114] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147–1163.

[115] SIVIC J, ZISSERMAN A. Video Google: a text retrieval approach to object matching in videos[C]//IEEE International Conference on Computer Vision (CVPR). 2003: 1470–1477.

[116] NISTER D, STEWENIUS H. Scalable recognition with a vocabulary tree[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2006: 2161–2168.

[117] SATTLER T, HAVLENA M, RADENOVIC F, et al. Hyperpoints and fine vocabularies for large-scale location recognition[C]//IEEE International Conference on Computer Vision (ICCV). 2015: 2102–2110.

[118] SCHÖNBERGER J L, PRICE T, SATTLER T, et al. A vote-and-verify strategy for fast spatial verification in image retrieval[C]//Asian Conference on Computer Vision (ACCV). 2017: 321–337.

[119] GALVEZ-LÓPEZ D, TARDÓS J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188–1197.

[120] SHEN T, ZHU S, FANG T, et al. Graph-based consistent matching for structure-from-motion[C]//European Conference on Computer Vision (ECCV). 2016: 139–155.

[121] SCHNABEL R, WAHL R, KLEIN R. Efficient RANSAC for point cloud shape detection[J]. Computer Graphics Forum, 2007, 26(2): 214–226.

[122] GONZÁLEZ D, PÉREZ J, MILANÉS V, et al. A review of motion planning techniques for automated vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(4): 1135–1145.

[123] VALENCIA R, ANDRADE-CETTO J, PORTA J M. Path planning in belief space with pose SLAM[C]//IEEE International Conference on Robotics and Automation (ICRA). 2011: 78–83.

[124] WERMELINGER M, FANKHAUSER P, DIETHELM R, et al. Navigation planning for legged robots in challenging terrain[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016: 1184–1189.

[125] LEE D T. Medial axis transformation of a planar shape[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1982, 4(4): 363–369.

[126] BOYKOV Y, VEKSLER O, ZABIH R. Fast approximate energy minimization via graph cuts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(11): 1222–1239.

[127] BOYKOV Y, KOLMOGOROV V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(9): 1124–1137.

[128] KOLMOGOROV V, ZABIN R. What energy functions can be minimized via graph cuts?[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26 (2): 147–159.

[129] GAO X S, HOU X R, TANG J, et al. Complete solution classification for the perspective-three-point problem[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(8): 930–943.

[130] KE T, ROUMELIOTIS S I. An efficient algebraic solution to the perspective-three-point problem[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4618–4626.

[131] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: An accurate O(n) solution to the PnP problem[J]. International Journal of Computer Vision, 2008, 81(2): 155–166.

[132] CUI H, SHEN S, GAO X, et al. Batched incremental structure-from-motion[C]// International Conference on 3D Vision (3DV). 2017: 205–214.

[133] MOULON P, MONASSE P. Unordered feature tracking made fast and easy[C]// European Conference on Visual Media Production (CVMP). 2012: 1.

[134] CASTALDO F, ZAMIR A, ANGST R, et al. Semantic cross-view matching[C]// IEEE International Conference on Computer Vision Workshop (ICCVW). 2015: 1044–1052.

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历

高翔，男，山东省莒南县人，1989.03.25 出生，中国科学院自动化研究所博士研究生。

2008.09-2012.06　中国海洋大学自动化专业，获工学学士学位

2012.09-2015.06　中国海洋大学控制理论与控制工程专业，获工学硕士学位

2015.09 至今　中国科学院自动化研究所模式识别与智能系统专业，攻读博士学位

已发表 (或正式接收) 的学术论文:

(1) Xiang Gao, Shuhan Shen, Yang Zhou, Hainan Cui, Lingjie Zhu, and Zhanyi Hu. "Ancient Chinese Architecture 3D Preservation by Merging Ground and Aerial Point Clouds," ISPRS Journal of Photogrammetry and Remote Sensing (P&RS), 2018.

(2) Xiang Gao, Lihua Hu, Hainan Cui, Shuhan Shen, and Zhanyi Hu. "Accurate and Efficient Ground-to-Aerial Model Alignment," Pattern Recognition (PR), 2018.

(3) Xiang Gao, Shuhan Shen, Zhanyi Hu, and Zhiheng Wang. "Ground and Aerial Meta-data Integration for Localization and Reconstruction: A Review," Pattern Recognition Letters (PRL), 2018.

(4) Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. "HSfM: Hybrid Structure-from-Motion," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

(5) Lingjie Zhu, Shuhan Shen, Xiang Gao, and Zhanyi Hu. "Large Scale Urban Scene Modeling from MVS Meshes," European Conference on Computer Vision (ECCV), 2018.

(6) Hainan Cui, Shuhan Shen, Xiang Gao, and Zhanyi Hu. "Batched Incremental Structure-from-Motion," International Conference on 3D Vision (3DV), 2017.

(7) Hainan Cui, Shuhan Shen, Xiang Gao, and Zhanyi Hu. "CSfM: Community-based Structure from Motion," IEEE International Conference on Image Processing (ICIP), 2017.

(8) Yang Zhou, Shuhan Shen, Xiang Gao, and Zhanyi Hu. "Accurate Mesh-based Alignment for Ground and Aerial Multi-view Stereo Models," IEEE International Conference on Image Processing (ICIP), 2017.

(9) Tianxin Shi, Shuhan Shen, Xiang Gao, and Lingjie Zhu. "Visual Localization Using Sparse Semantic 3D Map," IEEE International Conference on Image Processing (ICIP), 2019.

在审的学术论文:

(1) Xiang Gao, Shuhan Shen, Lingjie Zhu, Tianxin Shi, Zhiheng Wang, and Zhanyi Hu. "Complete Scene Reconstruction by Merging Images and Laser Scans," IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), major revision.

(2) Xiang Gao, Shuhan Shen, Lingjie Zhu, Haixia Wang, Hongmin Liu, and Zhanyi Hu. "Complete and Accurate Indoor Scene Capturing and Reconstruction Using a Mini Drone and a Ground Robot," Pattern Recognition (PR), under review.

(3) Xiang Gao, Hainan Cui, Lingjie Zhu, Tianxin Shi, and Shuhan Shen. "Multi-source Data Based 3D Digital Preservation of Large-scale Ancient Chinese Architecture: A Case Report," Virtual Reality and Intelligent Hardware (VRIH), under review.

致 谢

在自动化所的学习与生活令人难忘，值此论文完成之际，谨向四年间关心与帮助过我的所有人表示感谢！

感谢我的导师胡占义研究员。是您通过敏锐的学术洞察力为我指明了研究方向。在组会报告上您的每一个建议，在论文修改中您的每一条批注，都让我获益匪浅。您广博的专业知识，严谨的治学风范以及敬业的工作态度，无时无刻不深深地感染与激励着我。您对我的悉心指导、耐心鼓励和严厉批评，都是我今后工作与生活中的宝贵财富，我将铭记于心！

感谢我的导师申抒含副研究员。您扎实的学术基础，渊博的专业知识以及丰富的科研经验为我论文的完成提供了巨大的帮助。是您与我的一次次悉心讨论使我在陷入迷茫时激发出了新的研究灵感，是您对我的一次次耐心鼓励让我在遇到困难时拥有了继续研究的动力。您作为科研人员，学生导师与授课教师都是我需要学习的榜样！

感谢崔海楠老师。您亦师亦友，是我科研道路上的领路人，科研工作中的合作伙伴。您对所在专业领域的深入的了解，丰富的经验与独到的见解都让我深受启发！

感谢机器视觉课题组的其他各位老师，老师们对我的帮助与指导，给我的建议与意见我都会牢牢记住！

感谢机器视觉课题组的各位同窗：朱灵杰，周洋，时天欣，赵飞，刘养东等。和你们的讨论与交流使我收获颇丰。你们对我在科研上的帮助与在生活中的陪伴都是我美好的回忆！

感谢我的家人与朋友对我多年求学生涯的支持以及在我背后的默默付出与奉献！

2019 年 6 月于北京