第 43 卷 第 24 期/2023 年 12 月/光学学报 研究论文

光学学报

一种鲁棒的月面宽基线图像特征匹配方法

彭齐浩[1]，赵腾起[1]，刘传凯[2,3]，项志宇[1,4*]

[1]浙江大学信息与电子工程学院，浙江 杭州 310027；

[2]北京航天飞行控制中心，北京 100190；

[3]航天飞行动力学技术重点实验室，北京 100190；

[4]浙江省信息处理与通信网络重点实验室，浙江 杭州 310027

摘要 针对目前图像匹配算法在月面宽基线、弱纹理和光照变化等条件下匹配成功率低的问题，提出基于视图合成与全局注意力的月面图像匹配方法。首先对同站点月面双目图像使用稀疏视差虚真值训练立体匹配网络，完成同站点图像的三维重建。基于场景深度，结合站点之间惯导先验位姿将待匹配图像转为新的合成视图用于匹配，解决不同站点宽基线图像对之间图像重叠度低、视角变化大等问题。进一步使用基于Transformer的图像匹配网络，提高弱纹理场景下的图像匹配性能，并在后处理阶段引入考虑平面退化的外点滤除方法。在真实月面宽基线图像数据集的结果表明，相比现有算法，提出的匹配算法大幅度提高了宽基线场景下的月面图像匹配精度与成功率，为月球车大跨度行驶中的自主视觉定位提供了重要基础。

关键词 图像处理；月面图像匹配；特征提取；视图合成；三维重建

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS230498

1 引 言

月球是世界航天强国竞相探索与开采的热点领域。我国先后发射了嫦娥三号和四号探测器，并利用“玉兔号”和“玉兔2号”月球车开展了月面较大范围的长周期巡视。为了最大化月球车每次的行驶距离、提升遥操作模式下的探测效率，“玉兔”月球车采用大间距行进模式，每次行进距离约为6～10 m。这使得相邻导航站点距离较大，拍摄的图像存在较大的旋转、平移和尺度等宽基线变化，同时图像重叠度低且区域形态差异大，再加上月面光照变化、纹理较弱等特点，为站点间的图像特征匹配带来了极大的困难。目前“玉兔”月球车通过惯导与视觉相联合的方法来完成定位[1]，将惯导获得的定位结果作为初始位姿，通过前后站点图像人工筛选的视觉匹配以及图像特征点的三维空间关系来获取月球车最终的位姿。不同站点图像特征的精确匹配是视觉定位的难点，也是其中最关键的一个步骤。基于Affine-SIFT[2]的匹配算法已成功应用到前后站高缩放高形变图像的匹配中[1]，但是受限于表观特征匹配的局限性，依然存在大量的误匹配，需要人工进行辅助筛选或点取正确匹配，很大程度上影响地面遥操作处理效率。因此，提升不同站点图像特征匹配的鲁棒性以实现视觉定位的自动化，是迫切需要解决的问题。

目前，图像特征提取匹配算法可以分为两类：基于人工设计的传统特征提取匹配算法和基于深度学习的方法。基于人工设计的传统特征提取匹配算法比较成熟，文献[3]提出了一种尺寸不变性特征变化（SIFT）算法，通过构建高斯差分多尺度金字塔（DOG）提取具有尺度不变性的特征点，同时保证提取的特征具有一定的稳定性和抗噪性。针对图像特征提取算法实时性不高的问题，文献[4]在SIFT基础上提出了加速稳健特征（SURF）算法，使用积分图与合成滤波的方式加快特征点检测。文献[5]在构建描述子时利用主成分分析降维，提升了匹配速度，但是匹配性能略有下降。后续于子雯等[6]和苗延超等[7]分别在SIFT基础上提出了改进版SIFT算法与OS-SIFT算法解决异源图像匹配问题。为解决大视角场景下图像特征匹配问题，文献[2]在SIFT算法基础上提出了基于仿射不变性的匹配（ASIFT）算法，该算法基于物体局部可以近似为平面且透视效应可以忽略不计的假设，利用空间投影变换模型模拟视角变换，取得了不错的匹配效果，但在处理月面宽基线图像匹配时效果仍然欠佳。

近年来深度学习发展迅速，并在特征匹配领域取

收稿日期：2023-02-03；**修回日期**：2023-03-02；**录用日期**：2023-03-12；**网络首发日期**：2023-05-08

基金项目：航天飞行动力学技术重点实验室基金（2022-JYAPAF-F1027）

通信作者：*xiangzy@zju.edu.cn

得了很大进展。基于深度学习的特征提取匹配算法根据是否需要检测特征点又分为基于检测器(detector-based)与无需检测器(detector-free)两类。Detector-based的方法通过卷积神经网络检测特征点并生成对应的描述子,典型的方式有SuperPoint[8]、D2Net[9]、R2D2[10]和ASLFeat[11],这些方法的共同之处是均使用一个编码器对图像进行特征处理,生成高维的特征向量,之后通过一个或多个解码器检测特征点并生成对应的描述子。相比detector-based的方法,detector-free的方法省去了从图像中提取特征点并生成描述子的步骤,直接从图像对中得到像素的匹配关系。由于不需要检测特征点,detector-free的方法对弱纹理、重复纹理、光照变化等场景具有更好的匹配结果。Detector-free的思想最早可以追溯到SIFT Flow[12],它通过对比性损失函数为每个像素生成描述子,并通过最近邻的方式进行特征匹配。NCNet[13]和Sparse NCNet[14]通过端到端的方式直接提取致密的匹配,首先通过构建4D匹配代价体来编码所有可能的匹配,之后通过4D卷积惩罚匹配代价体保证所有匹配的邻居具有一致性关系。SuperGlue[15]是一种能够同时进行特征匹配和滤除外点的网络,其输入为两张图像对应的特征点及其描述子,通过图神经网络(GNN)建立特征点的对应关系,输出两张图像特征点的匹配关系,之后刘磊等[16]将其应用到合成孔径雷达图像与可见光图像的配准之中。LoFTR[17]提出由粗到细的匹配思想,首先在低分辨率的特征图上找到粗匹配,之后根据粗匹配在高分辨的特征图上预测一个匹配偏移量,最终得到亚像素级别的匹配关系。与使用匹配代价体搜索对应关系的稠密匹配方法相比,该方法使用Transformer[18]中自注意力和交叉注意力机制获取全图的感受野,能够在低纹理区域产生密集匹配,对弱纹理场景有很好的适用性。Tang等[19]提出四叉树注意力,将Transformer中注意力机制的运算复杂度从二次复杂度降低到了线性复杂度,在LoFTR上的应用也提高了图像特征匹配的精度。

月面环境是一种特殊的极端环境,其光照变化剧烈、纹理较为单一匮乏,前后站点月面图像之间存在大尺度旋转平移变化,现有的图像匹配算法依然很难实现月面(宽基线)图像的自动匹配。本文受弱纹理下鲁棒匹配算法LoFTR的启发,提出了一种适用于月面宽基线图像的匹配方法DepthWarp-LoFTR。采用惯导先验位姿与双目深度构建合成新的视图作为匹配桥梁,减轻待匹配图像的旋转和尺度变化,在此基础上利用LoFTR网络完成合成视图与另一站点图像的匹配,并在外点滤除阶段考虑基础矩阵平面退化情况。最后将匹配结果返回到原始宽基线图像对。基于月面实际图像的实验结果表明,该方法有效提高了特征匹配的成功率。

2 月面宽基线图像匹配算法

2.1 算法总体框架

本文的月面宽基线图像匹配算法如图1所示。针对前后站月面图像I_1与I_2,基于立体匹配网络GwcNet对同站点校正之后的双目图像进行立体匹配,结合视差与深度关系,将视差图转换成深度图,结合前后站惯导位姿合成新视图$\hat{I}_1$,利用基于Transformer的图像匹配算法LoFTR将新合成图像与后一站点图像进行匹配,经过外点滤除之后得到匹配结果,将该结果映射到原图坐标系下得到最终匹配关系。

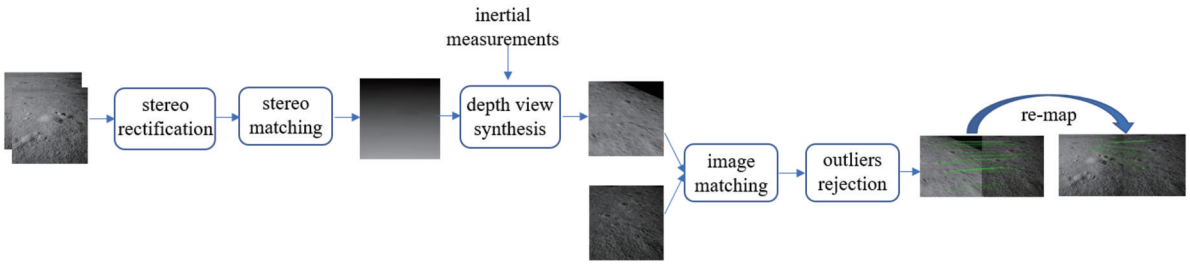

图1 月面宽基线图像匹配算法DepthWarp-LoFTR

Fig. 1 Matching algorithm for wide-baseline lunar images DepthWarp-LoFTR

2.2 月面同站点双目重建

采用GwcNet[20]网络进行月面同站点图像的立体匹配。为了提高月面场景下视差估计的精度,本文在月面场景下对GwcNet网络进行模型精调,在月面图像较少的情况下提高网络在月面场景下的泛化能力。采用左右目特征点匹配的方式生成稀疏视差伪真值,使用该伪真值训练GwcNet网络。对于校正之后的左右目图像,分别提取SIFT特征点,通过K近邻方式得到高质量匹配。为了滤除可能存在的误匹配,判断左右目匹配点的纵坐标之差,当纵坐标差的绝对值小于5个像素时,认为是正确的匹配,即$//p(y)-p^*(y)//<5$,其中p与p^*为左右目匹配点纵坐标,相比采用随机抽样一致(RANSAC)算法进行误匹配滤除能够更多地保留远处匹配点,从而保留远处的稀疏视差真值。当获取左右目匹配结果之后,视差为左右目匹配点横坐标之差。

2.3　基于场景深度的视图合成

月面前后站点距离远，拍摄的图像复合了较大的旋转平移以及尺度变化，导致图像重叠度小，肉眼很难分辨对应区域，给匹配算法带来了巨大的困难。本文基于同站点双目立体匹配估计的场景深度，结合前后站点之间的先验惯导位姿进行视图合成，减轻宽基线图像匹配的困难。基于场景深度的视图合成算法如图2所示，输入为前一站图像、后一站点深度图、后一站点到前一站点的惯导位姿变化($\boldsymbol{R}$,$\boldsymbol{t}$)，输出为合成的前一站点视图。

设前后站点图像分别为I_1、I_2，合成图像为$\hat{I}_1$，$\boldsymbol{p}_2$为后一站点图像的齐次像素坐标，$\boldsymbol{K}$为相机内参，D_2为后一站图像深度图，$T_{2\to1}$为后一站点到前一站点的惯导位姿变换，那么$\boldsymbol{p}_2$在前一站点像素坐标系下的投影坐标$\boldsymbol{p}_1$的计算如下所示：

$$p_1 \sim K T_{2\to1} D_2(p_2) K^{-1} p_2。\tag{1}$$

由于$\boldsymbol{p}_1$一般为浮点数，可以通过双线性插值的方式得到合成图像在$\boldsymbol{p}_2$位置处的灰度值$\hat{I}_1(\boldsymbol{p}_2)$，如下式所示：

$$\hat{I}_1(p_2)=I_1(p_1)=\sum_{i\in\{t,b\},j\in\{l,r\}} w^{ij} I_1(p_1^{ij}),\tag{2}$$

式中，w^{ij}表示与$\boldsymbol{p}_1$最近的四个点权重，其与$\boldsymbol{p}_1$和$\boldsymbol{p}_1^{ij}$的距离成线性比例关系。

2.4　弱纹理特征匹配网络

在月面弱纹理和光照变化条件下，一般的检测特征点并生成描述子的方法很难得到足够多的正确匹配。本文采用基于Transformer的直接预测匹配关系的网络LoFTR。LoFTR是一种detector-free的图像匹配方法，不需要从图像中提取特征点，直接从两幅图像中找点像素级别的2d-2d匹配点对，相比一般的基于卷积神经网络(CNN)的深度网络，LoFTR在使用CNN提取特征图之后，又通过Transformer模块引入全局感受野，每个特征位置除了聚合了局部的特征信息之外，更与全局的特征建立关联，弱纹理场景下匹配效果比detector-based的方法鲁棒性更好。网络采用由粗到细的逐级匹配方式，首先在粗粒度上建立像素的密集匹配，之后在细粒度上对粗匹配进行修正，得到更精细的匹配。与使用匹配代价空间搜索匹配关系的稠密匹配方法相比，LoFTR模拟人眼在搜寻匹配时会借助局部区域上下文信息的特点，使用Transformer中的自注意力与交叉注意力层来获取两张图像的特征描述，Transformer提供的全局感受野使得LoFTR能够在低纹理以及重复纹理区域获取大量致密的匹配。整体网络结构如图3所示，主要由四部分组成：

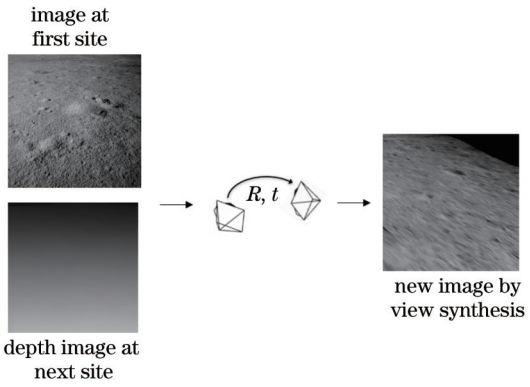

图2　基于场景深度的视图合成

Fig. 2　View synthesis based on scene depth

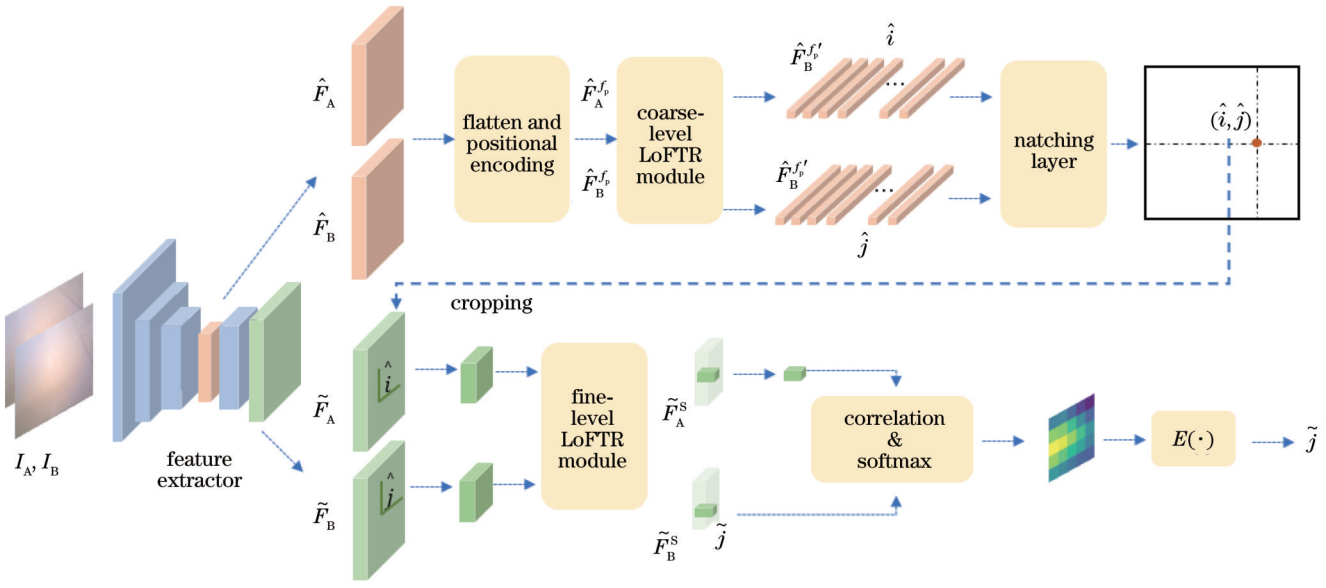

图3　LoFTR网络

Fig. 3　LoFTR network

1)特征提取模块。该模块使 ResNetFPN[21]作为特征提取器分别从图像 I_A 与图像 I_B 中提取多尺度的特征图，粗粒度特征图为 $\hat{F}_A$、$\hat{F}_B$，大小为 $\left(\frac{W}{8}\right)\times\left(\frac{H}{8}\right)\times 256$，细粒度特征度为 $\tilde{F}_A$、$\tilde{F}_B$，大小为 $\left(\frac{W}{2}\right)\times\left(\frac{H}{2}\right)\times 128$，其中 W、H 分别为原图的长与宽。

2)将得到的粗粒度特征图展成一维向量，并通过位置编码添加位置信息，输入 LoFTR 模块。LoFTR 模块由自注意力和交叉注意力模块组成，注意力层输入为查询向量($\boldsymbol{Q}$)、键向量($\boldsymbol{K}$)以及值向量($\boldsymbol{V}$)。查询向量与键向量点积之后作为权重，从值向量中检索信息，计算公式为 $\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\right)\boldsymbol{V}$。自注意力模块输入为同一幅图像对应的特征图，交叉注意力模块输入为两张图对应的特征图。经过自注意力与交叉注意力模块之后得到增强后的特征。

3)粗匹配模块。对于 LoFTR 模块增强的特征使用点积计算所有位置处的匹配矩阵，之后通过 dual-softmax 的方式计算最优匹配，通过相互最近邻剔除误匹配，并筛选置信度大于指定阈值的匹配形成粗匹配集合 $\boldsymbol{M}_c$。

4)对于粗匹配 $\boldsymbol{M}_c$ 中的一个匹配 $\left(\hat{i},\hat{j}\right)$，从细粒度特征图 $\tilde{F}_A$ 和 $\tilde{F}_B$ 中以 $\hat{i}$ 与 $\hat{j}$ 为中心裁剪出局部窗口大小为 $w\times w$(实验中 w 取 5)的特征图，经过细匹配阶段的 LoFTR 模块进行特征增强，提取精细匹配特征，记作 $\tilde{F}_A^s$ 与 $\tilde{F}_B^s$，计算 $\tilde{F}_A^s$ 中心向量与 $\tilde{F}_B^s$ 的匹配概率，之后计算概率分布，得到 $\tilde{F}_B^s$ 中亚像素精度的匹配位置。

LoFTR 网络使用 MegaDepth 数据集[22]的匹配真值进行训练，匹配真值通过图像对应的深度图和位姿计算得出。

2.5 外点滤除

对于月面图像，匹配点大多数位于较平坦的月面上，通过计算基础矩阵 F 来进行外点滤除会遇到平面情况 F 矩阵退化的情况。考虑平面退化情况，本文使用改进的随机抽样一致性算法 DGENSAC[23]算法进行外点滤除。过程如下：随机选取七对匹配点计算 F 矩阵，当有五组或以上的匹配点位于一个平面上时，考虑平面退化情况，额外计算单应性矩阵 $\boldsymbol{H}$，根据平面-视差算法[24]通过 $\boldsymbol{H}$ 矩阵外加额外两对匹配点计算 F 矩阵模型，进行匹配外点滤除。

3 实验结果与分析

3.1 数据集

本文使用“玉兔 2 号”月球车采集的真实月面图像用于算法评估，简称 Moon 数据集。数据集主要由两部分组成：第一部分包括“玉兔 2 号”月球车在 2019 年 7 月到 8 月之间采集的五个连续站点的总共 55 对双目图像，图像总数为 110，用于训练双目重建；第二部分包括 12 组前后站宽基线图像用于测试宽基线匹配效果，每一组包含 4 张图像，由前后站点的双目图像组成，图像总数为 48。其中 12 组宽基线图像匹配对的位姿真值由地面遥操作中心的“视觉-惯性定位算法[1]”计算得出，通过位置和旋转角的方式进行保存，实际使用时转换成矩阵表达形式，作为位姿真值衡量前后站宽基线图像匹配关系的好坏。

3.2 同站点双目重建以及视图合成

1)实验设置与评价指标

本文对 Moon 数据集中 5 个连续站点的 55 对双目图像生成稀疏视差真值，训练 GwcNet 网络，稀疏视差真值中有效视差占总体像素的 1%～2%。网络加载 sceneflow 数据集[25]预训练权重，使用 ADAM 优化器，超参数 $\beta_1=0.9$，$\beta_2=0.999$，图像分辨率为 1024×1024，批大小为 1，训练轮次为 500，初始学习率为 0.001，在第 200 轮和第 400 轮结束时学习率分别衰减到原始的 1/10。获得视差图之后，根据基线与焦距信息将视差转换成对应的深度。本文将获得的视差估计值与稀疏视差真值转换成深度后进行对比，来定量统计三维重建精度，如下式所示：

$$e_{\mathrm{rr}}=\frac{1}{N}\sum_{i=0}^{i=N}\left\|z_{\mathrm{est}}-z_{\mathrm{gt}}\right\|, \tag{3}$$

式中：z_{est} 与 z_{gt} 分别表示深度估计值与真值；N 表示真值视差点的个数。

2)重建结果

本文在 Moon 数据集上将精调的模型 GwcNet(Moon)与初始权重模型 GwcNet(sceneflow)以及传统的半全局立体匹配算法 SGBM 算法进行了定性与定量对比，其中 SGBM 由 OpenCV 自带函数 StereoSGBM_create 实现，视差数目设置为 224，块大小为 25。图 4 显示了三种不同算法视差估计的结果，颜色由黄变蓝表示视差值由大变小。如图 4(a)中区域 1～3 所示，传统的 SGBM 算法在图像边缘、月球坑边缘，以及重复性纹理较多的区域会出现视差估计错误或者视差空洞的情况。GwcNet(sceneflow)在图像边缘会出现视差估计错误的情况，如图 4(b)中区域 2～3 所示。月球坑边缘视差预测的细节也比较粗略，如图 4(b)中区域 1 所示。GwcNet(Moon)取得了最好的视差估计结果，在边缘处也保留了较好的视差细节。

本文在 12 组前后站图像上按式(3)定量统计了 3～7 m、8～12 m、13～19 m 不同距离范围内的三维重建精度，结果如表 1 所示，在月面场景下精调之后的立体匹配网络 GwcNet(Moon)取得了最好的重建精度，可以用于指导后续视图合成。

3)深度视图合成结果

图 5 第一行第二行显示了两组前后站图像，可以看出前后站图像存在较大的尺度、平移、旋转变化，从

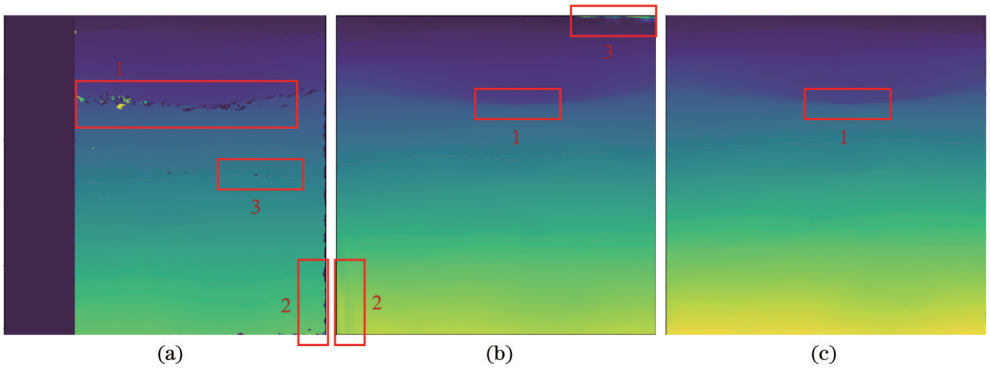

图 4　不同算法视差估计结果。(a)SGBM;(b)GwcNet(sceneflow);(c)GwcNet(Moon)

Fig. 4　Disparities of different algorithms. (a) SGBM; (b) GwcNet (sceneflow); (c) GwcNet (Moon)

表 1　三维重建精度

Table 1　Accuracy of 3D reconstruction

Algorithm	3D reconstruction error /m		
	3-7 m	8-12 m	13-19 m
SGBM	0.056	0.304	0.709
GwcNet (sceneflow)	0.012	0.073	0.225
GwcNet (Moon)	0.008	0.052	0.166

左到右三列分别为前一站图像、根据惯导初始位姿与深度图在后一站点的合成图像、后一站点图像。图中方形框标记了前一站图像、合成图像、后一站图像上某一块公共的区域,可以看出合成图像与后一站点图像中相同区域大小、位置基本一致,给后续的宽基线月面图像匹配奠定了基础。

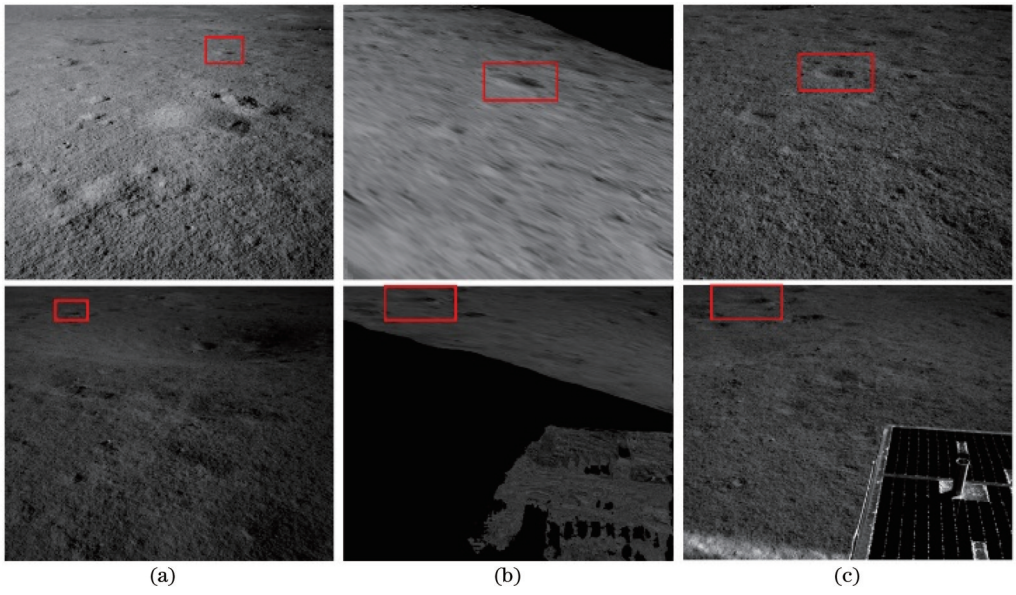

图 5　深度视图合成结果。(a)前一站点图像;(b)前一站点合成图像;(c)后一站点图像

Fig. 5　Results of depth view synthesis. (a) Image at previous site; (b) synthetic image at previous site; (c) image at next site

3.3　月面宽基线图像匹配

1)评价指标

为了定量评估前后站宽基线月面图像匹配的好坏,将 2d-2d 匹配关系通过 DEGENSAC 算法计算基础矩阵 ,从 F 矩阵中恢复出旋转与平移。由于平移向量缺少尺度信息,对于估计的旋转与平移向量,分别计算与位姿真值中旋转向量平移向量的角度误差(以°为单位),如下式所示:

$$\begin{cases} e_1 = \operatorname{acccos}\left(\dfrac{t \cdot t^*}{|t||t^*|}\right) \\ e_2 = \operatorname{acccos}\left(\dfrac{\theta \cdot \theta^*}{|\theta||\theta^*|}\right) \end{cases}, \tag{4}$$

式中:t 与 t^* 分别表示平移量的真值与估计值;θ 与 θ^* 分

别表示旋转量的真值与估计值，位姿真值通过由离线的“视觉-惯性定位算法”[1]计算而来。

2）定量与定性实验结果

本文选取 Moon 数据集中 12 对前后站宽基线图像对用于评估匹配算法 LoFTR 与 ASIFT。对于匹配网络 LoFTR，输入为合成的前一站图像与后一站点图像，输出为 2d-2d 匹配关系，通过 DEGENSAC 算法进行外点滤除。对于 ASIFT 算法，通过 K 近邻的方式寻找最优匹配和次优匹配，当最优匹配距离与次优匹配距离之比小于 0.9 时，保留最优匹配，最后通过 DEGENSAC 算法进行外点滤除。对于得到的匹配结果，通过对极几何恢复出前后站点的旋转和平移，并与来自“视觉-惯性定位算法”的位姿真值进行对比衡量匹配的好坏。表 2 显示了 LoFTR 与 ASIFT 算法在 Moon 数据集中 12 组前后站宽基线图像上的匹配结果，其中“Inliners/Matches”表示匹配内点数以及匹配数量，"x"表示匹配失败，没有对应的结果。对于 LoFTR 算法匹配数为网络直接输出的结果，匹配内点数为 DEGENSAC 算法滤除之后的结果。对于 ASIFT 匹配数为经过 K 近邻比率过滤之后的结果，匹配内点数为 DEGENSAC 算法滤除之后的结果。“r_{err}”与“t_{err}”分别表示旋转误差与平移误差，单位为°，由式（4）计算得出，最后一行“av”表示平均值。本文在 12 组前后站宽基线图像上进行了实验，当旋转平移平均误差小于 70°时，认为匹配成功。由于前后站复合了较大的旋转平移变化，再加上月面本身弱纹理、重复性纹理等特点，现有的图像匹配算法如 LoFTR 与 ASIFT 在前后站图像上匹配成功率很低，其中 LoFTR 算法在 12 组中仅有 4 组匹配成功，匹配成功率为 33.33%，ASIFT 算法匹配成功率为 16.67%。

图 6 显示了 LoFTR 与 ASIFT 算法在 2 号与 12 号前后站宽基线图像对上的匹配结果。由于前后站图像复合了大尺度的旋转平移变化，ASIFT 算法在月面重复性纹理场景下很难得到准确的匹配，其中 2 号图像对匹配失败，12 号图像对匹配成功，但与人工点取的匹配点对相比仍存在较多误匹配。相比 ASIFT 算法，基于 Transformer 的 LoFTR 算法在月面弱纹理、重复性纹理场景下取得了更好的匹配结果，对宽基线场景也具有更强的适用性，但依然存在匹配成功率不高以及匹配精度低的问题，对于匹配成功的图像依然会存在一定的误匹配。

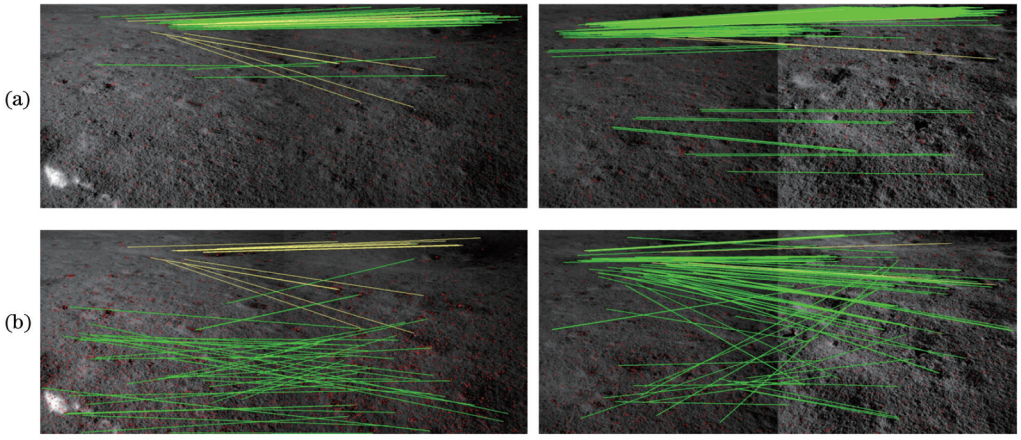

图 6 LoFTR 和 ASIFT 在 2 号、12 号前后站图像上的匹配结果。(a) LoFTR；(b) ASIFT

Fig. 6 Image matching results on No. 2 and No. 12 image pairs of LoFTR and ASIFT. (a) LoFTR; (b) ASIFT

表 2 LoFTR 和 ASIFT 算法匹配成功结果

Table 2 Successful matching results of LoFTR and ASIFT

No.	LoFTR			ASIFT		
	Inliners/Matches	r_{err}	t_{err}	Inliners/Matches	r_{err}	t_{err}
2	145/751	**28.73**	**35.14**		x	
4	64/447	**64.64**	**22.67**		x	
9	124/1009	**1.94**	**0.55**	875/5316	2.54	1.56
12	211/822	**2.82**	**3.21**	105/1045	13.73	15.34
av	—	24.53	15.39	—	8.14	8.45

对于 DepthWarp-LoFTR 算法，视图合成部分使用了月面图像的惯导位姿和立体深度信息，匹配采用网络 LoFTR，通过 DEGENSAC 算法进行外点滤除，最后将匹配转换到原图像中得到前后站宽基线图像的匹配。作为对比，本文使用 ASIFT 方法替换 LoFTR 网络，匹配阶段依然使用合成图像，记作 DepthWarp-

ASIFT 算法。对于得到的匹配结果，依然通过对极几何恢复出前后站点的旋转和平移，与位姿真值进行对比衡量匹配的好坏。表 3 显示了 DepthWarp-LoFTR 和 DepthWarp-ASIFT 算法在 Moon 数据集中 12 组前后站宽基线图像上的匹配结果。对比 ASIFT 算法，利用合成图像作为匹配桥梁的 DepthWarp-ASIFT 匹配成功 5 对，匹配成功率为 41.67%，算法匹配成功率更高，同时取得了更高的匹配精度，证明了视图合成方案在前后站宽基线月面图像匹配的有效性。最后，DepthWarp-LoFTR 算法在 12 对宽基线图像对中匹配成功 10 对，匹配成功率为 83.33%，相比 LoFTR 大幅度提升了匹配成功率与匹配精度，且大部分情况下其匹配精度要优于 DepthWarp-ASIFT 算法。由于个别序号图像对匹配误差较大，如表 3 中序号 4 与序号 7，导致 DepthWarp-LoFTR 算法平移误差平均值要略大一些。

图 7 显示了 1 号、5 号前后站点图像匹配结果，其中 ASIFT 和 LoFTR 算法由于前后站图像复合了较大的旋转平移尺度变化，匹配失败。DepthWarp-ASIFT 在 1 号图像对上匹配成功，5 号图像对前后站图像表面变化很大，再加上月球车本体的干扰，导致匹配失败。相比之下，DepthWarp-LoFTR 获得了一致性更强同时更加致密的匹配结果。

表 3 DepthWarp-LoFTR 和 DepthWarp-ASIFT 的匹配结果

Table 3 Matching results of DepthWarp-LoFTR and DepthWarp-ASIFT

No.	DepthWarp-LoFTR			DepthWarp-ASIFT		
	Inliners/Matches	r_{err}	t_{err}	Inliners/Matches	r_{err}	t_{err}
1	454/765	**1.91**	**3.42**	71/414	67.36	38.11
2	217/660	34.96	31.07	82/301	**12.04**	**4.64**
3	284/466	**5.48**	**6.38**		x	
4	20/133	**44.70**	**80.13**		x	
5	230/480	**2.66**	**20.31**		x	
6	107/278	**6.53**	**9.71**		x	
7	106/308	**28.70**	**48.08**		x	
8		x			x	
9	665/1025	1.47	**0.69**	2378/5129	**0.52**	7.44
10	53/281	**9.36**	22.40	30/121	71.51	**14.24**
11		x			x	
12	1161/1487	**0.62**	**0.58**	184/297	3.81	4.05
av	—	13.64	22.28	—	31.05	13.70

图 8 显示了 DepthWarp-LoFTR 匹配失败的情况。8 号图像对为月球车在前后站点分别朝向前与朝向后拍摄的图像，图像之间光照变化明显，同一个区域（如图中的石块部分）在两张图中成像相差巨大，导致匹配失败。11 号图像对则是由于前一站点拍摄图像出现过曝和较大面积的阴影部分，导致匹配失败。

3）时间效率

本文将立体匹配网络 GwcNet（Moon）与特征匹配网络 LoFTR 均放在 CPU［Intel（R） Xeon（R） Silver 4210R CPU］上进行推理，统计算法耗时。用于视差预测的立体匹配由于包含了较多的 3D 卷积，其在 CPU 上推理时间较慢，LoFTR 网络由于内部使用了 Transformer 全局注意力的学习，其推理时间也相对较慢，各个模块的具体耗时如表 4 所示。DepthWarp-LoFTR 算法总耗时约为 39 s，能满足“玉兔 2 号”月球车地面遥操作系统对时间的要求。

4 结 论

本文提出了一种鲁棒的月面宽基线图像匹配算法 DepthWarp-LoFTR。对同站点双目图像，利用左右目特征点匹配生成稀疏视差，作为伪真值训练 GwcNet 网络进行致密立体匹配，实现同站点图像的三维重建。

表 4 各算法的运行时间

Table 4 Running time of different algorithms unit: s

Algorithm	Disparity prediction	Running time			Total
		View synthesis	Extraction and matching	Outlier rejection	
DepthWarp-ASIFT	23.295	4.710	8.312	0.487	36.804
DepthWarp-LoFTR	23.295	4.710	10.844	0.211	39.060

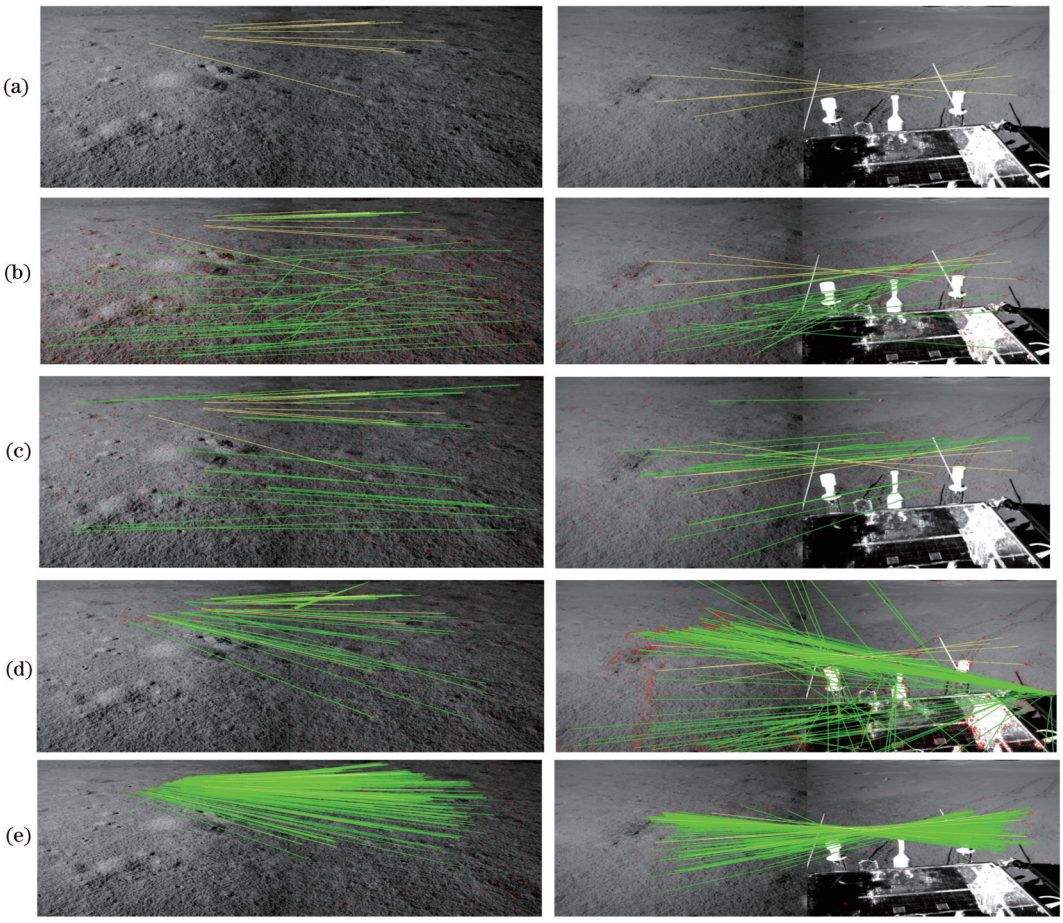

图 7　1 号、5 号前后站图像匹配结果。(a)真值;(b)ASIFT;(c)LoFTR;(d)DepthWarp-ASIFT;(e)DepthWarp-LoFTR

Fig. 7　Image matching results on No. 1 and No. 5 image pairs. (a) Ground-truth matches; (b) ASIFT; (c) LoFTR; (d) DepthWarp-ASIFT; (e) DepthWarp-LoFTR

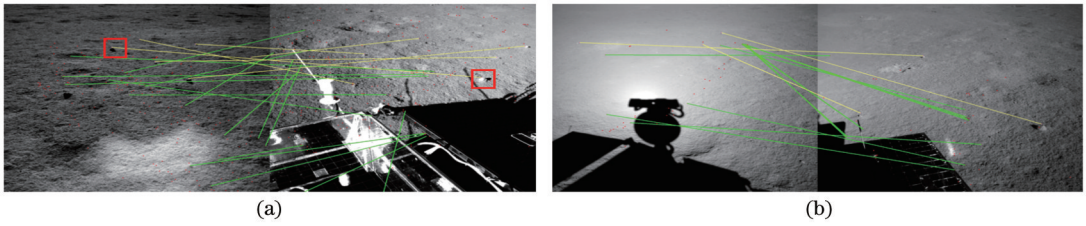

图 8　DepthWarp-LoFTR 匹配失败例子。(a)8 号;(b)11 号

Fig. 8　Matching failure cases of DepthWarp-LoFTR. (a) No. 8; (b) No. 11

结合惯导先验位姿,提出基于场景深度的视图合成算法,使用合成的前站点图像与后站点图像进行匹配,减轻前后站图像宽基线带来的特征匹配困难。在特征匹配阶段采用基于 Transformer 的 LoFTR 网络,大幅度提高了自动匹配的成功率与精度。在实际月面数据集上的实验结果表明,本文算法大幅提高了月面宽基线复杂场景下特征匹配的成功率,为“玉兔 2 号”月球车及后续我国探月四期中的月球车常态化巡视的遥操作自动视觉定位打下了坚实的基础。

参　考　文　献

[1] 刘传凯, 王保丰, 王镓, 等. 嫦娥三号巡视器的惯导与视觉组合定姿定位[J]. 飞行器测控学报, 2014, 33(3): 250-257.
Liu C K, Wang B F, Wang J, et al. Integrated INS and vision-based orientation determination and positioning of CE-3 lunar rover[J]. Journal of Spacecraft TT&C Technology, 2014, 33(3): 250-257.

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

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

[4] Bay H, Tuytelaars T, Van Gool L. SURF: speeded up robust features[M]//Leonardis A, Bischof H, Pinz A. Computer vision-ECCV 2006. Lecture notes in computer science. Heidelberg: Springer, 2006, 3951: 404-417.

[5] Ke Y, Sukthankar R. PCA-SIFT: a more distinctive representation for local image descriptors[C]//Proceedings of the 2004 IEEE Computer Society Conference on Computer

Vision and Pattern Recognition, 2004. CVPR, June 27-July 2, 2004, Washington, DC, USA. New York: IEEE Press, 2004.

[6] 于子雯, 张宁, 潘越, 等. 基于改进的SIFT算法的异源图像匹配[J]. 激光与光电子学进展, 2022, 59(12): 1211002.
Yu Z W, Zhang N, Pan Y, et al. Heterogeneous image matching based on improved SIFT algorithm[J]. Laser & Optoelectronics Progress, 2022, 59(12): 1211002.

[7] 苗延超, 刘晶红, 刘成龙, 等. 基于改进OS-SIFT的可见光与SAR图像自动配准[J]. 激光与光电子学进展, 2022, 59(2): 0228006.
Miao Y C, Liu J H, Liu C L, et al. Automatic registration of optical and SAR images based on improved OS-SIFT[J]. Laser & Optoelectronics Progress, 2022, 59(2): 0228006.

[8] DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: self-supervised interest point detection and description[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 224-236.

[9] Dusmanu M, Rocco I, Pajdla T, et al. D2-net: a trainable CNN for joint description and detection of local features[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 8084-8093.

[10] Revaud J, Weinzaepfel P, de Souza C, et al. R2D2: repeatable and reliable detector and descriptor[EB/OL]. (2019-06-14)[2022-11-09]. https://arxiv.org/abs/1906.06195.

[11] Luo Z X, Zhou L, Bai X Y, et al. ASLFeat: learning local features of accurate shape and localization[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6588-6597.

[12] Liu C, Yuen J, Torralba A. SIFT flow: dense correspondence across scenes and its applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 978-994.

[13] Rocco I, Cimpoi M, Arandjelović R, et al. Neighbourhood consensus networks[EB/OL]. (2018-10-24)[2022-11-09]. https://arxiv.org/abs/1810.10510.

[14] Rocco I, Arandjelović R, Sivic J. Efficient neighbourhood consensus networks via submanifold sparse convolutions[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12354: 605-621.

[15] Sarlin P E, DeTone D, Malisiewicz T, et al. SuperGlue: learning feature matching with graph neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 4937-4946.

[16] 刘磊, 李元祥, 倪润生, 等. 基于卷积与图神经网络的合成孔径雷达与可见光图像配准[J]. 光学学报, 2022, 42(24): 2410002.
Liu L, Li Y X, Ni R S, et al. Synthetic aperture radar and optical images registration based on convolutional and graph neural networks[J]. Acta Optica Sinica, 2022, 42(24): 2410002.

[17] Sun J M, Shen Z H, Wang Y A, et al. LoFTR: detector-free local feature matching with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8918-8927.

[18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 6000-6010.

[19] Tang S T, Zhang J H, Zhu S Y, et al. QuadTree attention for vision transformers[EB/OL]. (2022-01-08)[2022-11-09]. https://arxiv.org/abs/2201.02767.

[20] Guo X Y, Yang K, Yang W K, et al. Group-wise correlation stereo network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 3268-3277.

[21] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[22] Li Z Q, Snavely N. MegaDepth: learning single-view depth prediction from Internet photos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2041-2050.

[23] Chum O, Werner T, Matas J. Two-view geometry estimation unaffected by a dominant plane[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-26, 2005, San Diego, CA, USA. New York: ACM Press, 2005: 772-779.

[24] Irani M, Anandan P. Parallax geometry of pairs of points for 3D scene analysis[M]//Buxton B, Cipolla R. Computer vision-ECCV '96. Lecture notes in computer Science. Heidelberg: Springer, 1996, 1064: 17-30.

[25] Mayer N, Ilg E, Häusser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4040-4048.

A Robust Feature Matching Method for Wide-Baseline Lunar Images

Peng Qihao[1], Zhao Tengqi[1], Liu Chuankai[2,3], Xiang Zhiyu[1,4*]

[1]*College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou* 310027, *Zhejiang, China*;

[2]*Beijing Aerospace Flight Control Center, Beijing* 100190, *China*;

[3]*National Key Laboratory of Science and Technology on Aerospace Flight Dynamics, Beijing* 100190, *China*;

[4]*Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Hangzhou* 310027, *Zhejiang, China*

Abstract

Objective The vision-based navigation and localization system of China's "Yutu" lunar rover is controlled by a ground teleoperation center. A large-spacing traveling mode with approximately 6-10 m per site is adopted for the rover to maximize the driving distance of the lunar rover and improve the efficiency of remote control exploration. This results in a significant distance between adjacent navigation sites, and considerable rotation, translation, and scale changes in the captured images. Furthermore, the low overlap between images and the vast differences in regional shapes, combined with weak texture and illumination variations on the lunar surface, pose challenges to image feature matching among different sites. Currently, the "Yutu" lunar rover employs inertial measurements and visual matches among different sites for navigation and positioning. The ground teleoperation center adopts inertial measurements as initial poses and optimizes the poses with visual matches by bundle adjustment to obtain the final rover poses. However, due to the wide baseline and significant surface changes of images at different sites, manual assistance is often required to filter or select the correct matches, significantly affecting the efficiency of the ground teleoperation center. Therefore, improving the robustness of image feature matching between different sites to achieve automatic visual positioning is an urgent problem to be addressed.

Methods To address the poor performance and low success rate of current image matching algorithms in wide-baseline lunar images with weak textures and illumination variations, we propose a global attention-based lunar image matching algorithm by the view synthesis. First, we utilize sparse feature matching methods to generate sparse pseudo-ground-truth disparities for the rectified stereo lunar images at the same site. Next, we finetune a stereo matching network with these disparities and perform 3D reconstruction for the lunar images at the same site. Then, we leverage inertial measurements among different sites to convert the original image into a new synthetic view for matching based on the scene depth, addressing the low overlap and large viewpoint changes among images of different sites. Additionally, we adopt a Transformer-based image matching network to improve matching performance in weak-texture scenes, and an outlier rejection method that considers plane degeneration in the post-processing stage. Finally, the matches are returned from the synthetic image to the original image, yielding the matches for wide-baseline lunar images at different sites.

Results and Discussions We conduct experiments on the real lunar dataset from the "Yutu 2" lunar rover (referred to as the Moon dataset), which includes two parts. The first part is stereo images from five continuous stations (employed for stereo reconstruction), and the second is 12 sets of wide-baseline lunar images from adjacent sites (for wide-baseline image matching testing). In terms of lunar 3D reconstruction, we calculate the reconstruction error within different distance ranges, where the reconstruction network GwcNet (Moon) yields the best reconstruction accuracy and reconstruction details, as shown in Table 1 and Fig. 4. Meanwhile, Fig. 5 illustrates the synthetic images obtained from the view synthesis scheme based on the inertial measurements between sites and the scene depth, which solves the large rotation, translation, and scale changes between adjacent sites. For wide-baseline image matching, existing algorithms such as LoFTR and ASIFT have matching success rates of 33.33% and 16.67% respectively as shown in Table 2. Our DepthWarp-LoFTR algorithm achieves a matching success rate of 83.33%, significantly improving the matching success rate and accuracy of wide-baseline lunar images (Table 3). Additionally, this algorithm increases the matching success rate from 16.67% to 41.67% compared to the ASIFT algorithm. We present the matching results of different algorithms in Fig. 7, where DepthWarp-LoFTR obtains more consistent and denser matching results compared to other methods.

Conclusions We propose a robust feature matching method DepthWarp-LoFTR for wide-baseline lunar images. For stereo images captured at the same site, the sparse disparities are generated through a sparse feature matching algorithm. These disparities serve as pseudo-ground truth to train the GwcNet network for 3D reconstruction of lunar images at the same site. To handle the wide baseline and low overlap of images from different sites, we propose a view synthesis algorithm based on scene depth and inertial prior poses. Image matching is performed on the synthesized current-site image

and the next-site image to reduce the feature matching difficulty. For the feature matching stage, we adopt a Transformer-based LoFTR network, which significantly improves the success rate and accuracy of automatic matching. Our experimental results on real lunar datasets demonstrate that the proposed algorithm greatly improves the success rate of feature matching in complex lunar wide-baseline scenes. This lays a solid foundation for automatic visual positioning of the "Yutu 2" lunar rover and subsequent routine patrols of lunar rovers in China′s fourth lunar exploration phase.

Key words image processing; lunar image matching; feature extraction; view synthesis; 3D reconstruction