基于模态转换和三维空间关系约束的多模高分遥感影像配准方法

李晓奎[1]，杨树文[1,2,3]，李子源[1]，王文举[1]，朱浩[1]，薛祎明[1]

1. 兰州交通大学 测绘与地理信息学院, 兰州 730070;
2. 地理国情监测技术应用国家地方联合工程研究中心, 兰州 730070;
3. 甘肃省测绘科学与技术重点实验室, 兰州 730070

摘　要：复杂城市场景中的密集建筑和显著高差，放大了多模态遥感影像间的非线性辐射误差和几何畸变，给高分辨率影像的高精度配准带来了挑战。现有配准方法多依赖于影像的纹理等二维特征，缺乏对三维空间信息的有效利用。为此，本文提出了一种基于模态转换和三维空间关系约束的多模态高分辨率遥感影像配准方法。首先，该方法使用跨模态影像翻译来减弱多模态影像间的辐射差异；其次，使用单目深度估计快速获取深度图，为后续的特征描述和匹配约束提供数据基础与空间先验；最后，提出了一种基于深度的三维空间关系约束匹配方法。该方法通过多特征图的特征点检测、深度联合描述符构建和三维空间关系约束来实现特征点的潜在检测、稳健描述和精确匹配。本文在4组代表性数据上，将所提方法与多种传统和深度学习方法进行了对比。实验结果表明，所提方法的平均NCM为510，较现有方法提升约0.92~5.22倍；平均RMSE为1.58像素，精度较现有方法提升约4.12~4.78倍。结果验证了本文方法在配准精度和综合性能上具有一定优势，能够有效抑制城市多模态影像中的非线性辐射差异和几何畸变，实现更高精度配准。本文使用的数据和相关资源可从以下地址获取：https://github.com/QAlana/CityMMReg-data。

关键词：图像配准，多模态，三维空间关系，城市复杂场景，深度图

中图分类号：TP701

引用格式：李晓奎,杨树文,李子源,王文举,朱浩,薛祎明.XXXX.基于模态转换和三维空间关系约束的多模高分遥感影像配准方法.遥感学报,XX(XX):1-15

LI Xiaokui, YANG Shuwen, LI Ziyuan, WANG Wenju, ZHU Hao, XUE Yiming. XXXX. Multimodal High-Resolution Remote Sensing Image Registration via Modality Translation and 3D Spatial Relationship Constraints. National Remote Sensing Bulletin, DOI:10.11834/jrs.20265468]

1　引言

高精度影像配准是开展影像融合、变化检测、目标识别与跟踪等高级应用的基础（Hou等，2024），而多模态影像的联合处理则能够进一步提升结果的可靠性、完整性与准确性（Wu等，2025）。例如，在灾害应急响应中，通过精确对齐灾后的合成孔径雷达和光学影像，可实现受损建筑物的全天候识别与风险评估（Sun等，2024）；在城市管理中，基于高精度配准的多时相影像变化检测，有助于实现建筑演变监测与土地利用变化分析（Zhou等，2024）。然而，在城市复杂场景下，高大地物分布密集、显著高差、遮挡和阴影效应显著增加了多模高分遥感影像高精度配准的难度（Zheng等，2024）。因此，针对城市多模态高分辨率遥感影像，开发鲁棒、高精度的配准方法具有重要的科学意义和现实应用价值。

目前，多模态遥感影像配准方法大都在二维空间中展开研究，这些方法主要可以分为两类：基于区域和基于特征的方法（杨等，2025；Zhang等，2025）。前者主要通过影像间的二维统计信息或结构相似性度量进行配准，例如基于信息论的

收稿日期：2025-11-17；**预印本：**XXXX-XX-XX

基金项目：国家自然科学基金(编号：42471428)

第一作者简介：李晓奎,研究方向为遥感影像智能处理。E-mail：lxk0921high@163.com

通信作者简介：杨树文,研究方向为遥感数字图像处理及信息自动提取、灾害遥感。E-mail：ysw040966@163.com

互信息（MI）（Wells等，1996）、基于结构特征的相位一致性方向直方图（HOPC）（Ye和Shen，2016）、方向梯度通道特征（CFOG）（Ye等，2019）、融合傅里叶和加权边缘密度的相位一致性方向直方图（FED-HOPC）（Ye等，2024）和快速秩局部自相似性（FRLSS）（Xiong等，2025）等方法。这类方法虽然能够处理多模态影像，但其依赖于大范围的强度或结构纹理信息，难以实现高精度和高鲁棒的匹配（叶等，2024）。相比之下，基于特征的方法侧重于通过特征点的稳健描述进行匹配，可在一定程度上减弱辐射差异带来的影响。例如，位置尺度定向不变特征变换（PSO-SIFT）（Ma等，2017）、辐射不变特征变换（RIFT）（Li等，2020）、绝对相位方向梯度直方图（HAPCG）（姚等，2021）、多方向张量索引特征图（MoTIF）（Yao等，2022）和位置-方向-尺度引导的几何与强度不变特征变换（POS-GIFT）（Hou等，2024）方法等。尽管这类方法在多模态影像匹配中更加稳健，但仍然将影像视作二维平面进行处理，忽略了真实场景中的三维空间关系。这类二维假设的方法难以准确刻画复杂城市场景下影像间的几何形变，导致二维纹理相似但三维空间位置不同的特征被错误匹配，从而限制了配准的精度与几何一致性。

针对影像二维假设的局限，一些研究尝试在配准过程中引入三维空间信息，以提升匹配的几何一致性和精度。例如，Kim等（2012）和Wang等（2022）利用遥感影像的姿态参数（如有理多项式系数）和数字高程模型对影像进行预处理，以减小影像间的旋转和尺度差异。然而，这类方法依赖额外的姿态参数和DEM数据，通常只能进行大范围的粗略校准，难以满足城市高分辨率影像精确配准的需求。Zheng等（2024）通过提取建筑物轮廓和阴影反演建筑高度，并在特征匹配阶段引入高度惩罚因子，降低阴影对纹理相似性的干扰。但是，这种方法依赖于可见光影像中的阴影信息，难以推广至多模态遥感影像场景。

基于以上分析，在城市复杂环境下，为实现多模态高分辨率遥感影像的高精度配准，需要着重关注以下三个方面：（1）跨模态差异带来的影响。由于传感器及其成像机制的不同，多模态影像在辐射响应、几何畸变和纹理表征等方面存在显著差异。（2）城市复杂场景引发的形变与遮挡。城市地区建筑分布密集且高差显著，在多视角成像和地物高度变化等因素的作用下，容易产生遮挡、阴影及视角畸变等复杂现象。（3）三维空间信息利用不足。针对城市场景特有的三维结构，现有的配准方法尚未对三维空间信息进行充分而有效的利用，从而限制了配准精度与鲁棒性的进一步提升。

鉴于此，本文提出了一种基于模态转化和三维空间关系约束的多模态高分辨率影像配准方法，以重点解决上述问题：（1）使用生成对抗网络进行模态转换，将高分辨率的城市跨模态影像映射至近似同模态的表征空间，有效缓解模态差异对特征匹配的干扰。（2）使用单目深度估计快速获取伪三维深度信息，并构建归一化深度联合描述符，以增强特征描述符在复杂几何形变条件下的鲁棒性和描述能力。（3）提出一种基于深度的三维空间关系约束机制，通过识别并剔除由深度不一致引起的伪匹配，弥补传统方法对三维信息利用不足的问题，提升城市复杂场景下的配准精度与稳定性。

2 研究方法

本文方法的整体框架如图1所示，主要包含以下几个关键步骤：首先，针对异源遥感影像间显著的辐射和纹理差异，使用pix2pixHD（Wang等，2018）模型进行跨模态图像翻译，将SAR和红外等影像转换为伪光学影像，以降低模态差异对后续特征提取和匹配的影响。其次，分别对参考光学影像和生成的伪光学影像进行单目深度估计，并通过深度对齐策略获取对齐深度图，为后续的三维几何约束提供数据支撑。最后，在特征匹配阶段，在多特征响应图上使用FAST算子（Rosten等，2010）获取候选特征点，构建融合深度信息的归一化联合特征描述符。同时，引入基于多尺度深度一致性校验的三维空间约束机制，剔除存在深度冲突或视差异常的匹配，实现复杂城市场景下多模态高分辨率遥感影像的稳健配准。

2.1 跨模态图像翻译

不同模态影像之间显著的非线性辐射差异，是多模态高分辨率遥感影像配准面临的首要挑战之一。传统基于区域和特征的方法难以在这些影像中建立可靠的对应关系（见图2）。需要说明的

是，为取得更清晰的可视化效果，本文在结果展示阶段将所有SAR和红外影像替换为对应的RGB影像，但特征提取和匹配仍然在异源影像上进行，该操作对配准结果无影响。为了缓解模态差异带来的影响，本文引入跨模态图像翻译技术，将SAR或红外影像映射至为与真实光学影像更为接近的伪光学影像（见图3），从而把异源影像配准问题转化为近似同源影像配准问题，也为后续的特征提取和匹配提供了更为友好的数据基础。Yu等（2024）的研究表明，生成对抗网络在跨模态图像翻译任务中具有较好的表现。

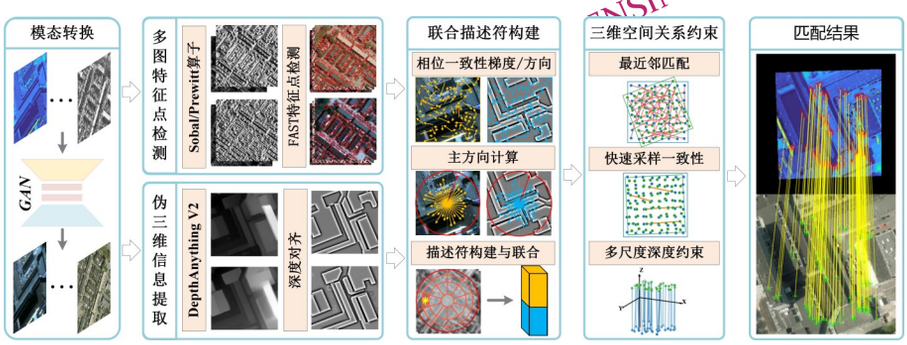


图1 所提方法流程

Fig. 1 Workflow of the proposed method

（a） MoTIF方法匹配结果和棋盘格图

（a） Matching results and checkerboard image of the MoTIF method

（b） EPC-SRC方法匹配结果和棋盘格图

（b） Matching results and checkerboard image of the EPC-SRCmethod

图2 现有方法的配准结果

Fig. 2 Registration results of existing methods

2.1.1 生成对抗网络概述

生成对抗网络（GAN）最初由Goodfellow等（2014）提出，其核心思想是通过生成器（Generator，G）和判别器（Discriminator，D）的对抗性训练，生成高质量的、与真实数据相似的虚假数据（Xia等，2023），它的基本原理如图3所示。在GAN模型的基础上，pix2pixHD通过引入由粗到细的生成器、多尺度判别器和改进的对抗损失，以生成高分辨率图像并添加更精细的纹理。pix2pixHD模型的生成器包含全局生成网络（G_1）和局部增强网络（G_2）两个子网络。其中G_1用于生成一张与输入图像一样大小的伪图像，G_2则是

为了输出更高分辨率图像，生成一张长高分别是输入图像长高两倍的伪图像。pix2pixHD的判别器则是在原图、原图的1/2和1/4降采样上，使用具有相同网络结构的鉴别器D1、D2、D3来分别判别这三个尺寸图像的真假。相应地，pix2pixHD模型的损失函数增加到了三个部分，分别是多尺度对抗损失、多尺度特征匹配损失和内容损失。总损失函数如下：

$$L_{pix2pixHD} = L_{GAN}(G,D) + \lambda_{FM}L_{FM}(G,D) + \lambda_{VGG}L_{VGG}(G) \quad (1)$$

其中，$L_{GAN}(G, D)$为对抗损失，$L_{FM}(G, D)$为特征匹配损失，$L_{VGG}(G)$为感知损失。λ_{FM}和λ_{VGG}为对应的权重系数，用于平衡不同的损失项，在原始论文中均取10。

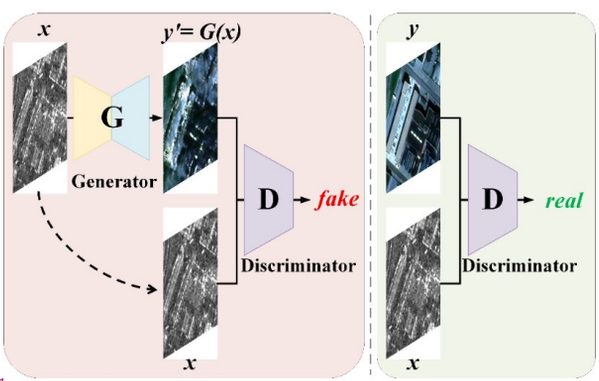


图3　图像翻译基础模型

Fig. 3　Basic model of image translation

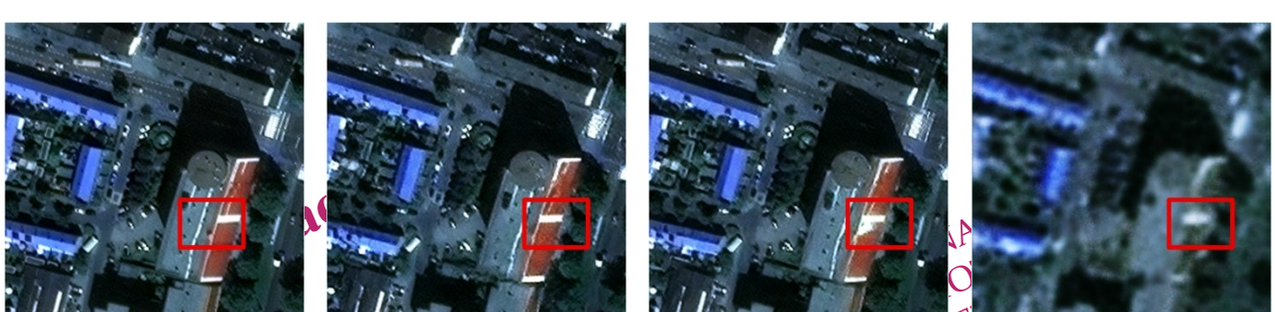
2.1.2　基于生成对抗网络的跨模态图像翻译

本节使用的训练数据来自DDHRNet数据集（Ren等，2022）和SpaceNet 6数据集（Shermeyer等，2020），从中筛选韩国浦项市和荷兰鹿特丹港的主城区影像并裁剪至256×256。表1是pix2pix（Isola等，2017）、pix2pixHD（G_1）和pix2pixHD（G_1+G_2）等模型在SpaceNet 6上的训练结果，图4为模型的实际转换效果。其中感知相似性（LPIPS）（Zhang等，2018）、峰值信噪比（PSNR）（Sara等，2019）和结构相似性指数（SSIM）（郑等，2025）是跨模态图像翻译领域常用的指标，用于评估生成图像的质量。PSNR的单位是分贝（dB），训练耗时（Time）的单位是小时（h），其余指标均为无纲量。由表1和图4可知，pix2pixHD（G_1+G_2）模型在图像生成指标上表现相对较优但提升幅度有限，且实际的转换效果不及pix2pixHD（G_1）。综合考虑下，本文在batchsize为8，epoch为200，其余超参数沿用Wang等（2018）默认值的条件下，对pix2pixHD（G_1）进行训练，实现SAR和IR影像向RGB影像的模态转换。

表1　三种模态转换模型的定量结果，最优结果用粗体标记

Table 1　Quantitative results of three modality translation models, with optimal results in bold

Method	LPIPS↓	SSIM↑	PSNR/dB↑	Time/h↓
pix2pixHD（G_1）	0.0676	0.7111	22.72	2.1
pix2pixHD（G_1+G_2）	**0.0639**	**0.7206**	**23.18**	17.1
pix2pix	0.5454	0.4648	18.93	**0.6**



（a）　目标影像　（b）　pix2pixHD（G_1结果）　（c）　pix2pixHD（G_1+G_2）结果　（d）　pix2pix结果

（a）　Target image　（b）　pix2pixHD（G_1）result　（c）　pix2pixHD（G_1+G_2）result　（d）　pix2pix result

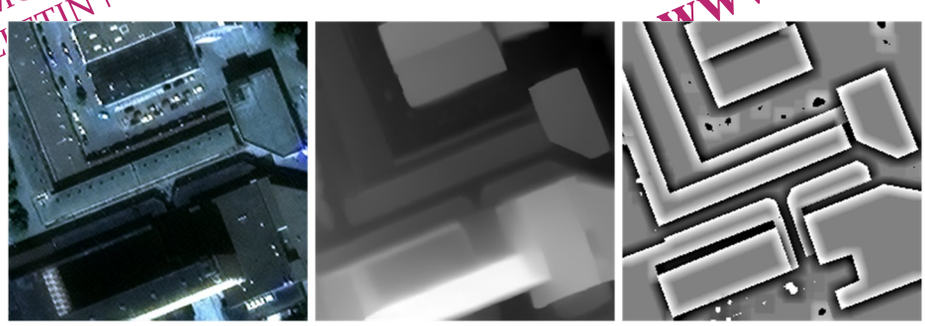
图4　三种模态转换模型定性结果

Fig. 4　Qualitative results of three modality translation models

2.2 基于单目深度估计的三维信息获取

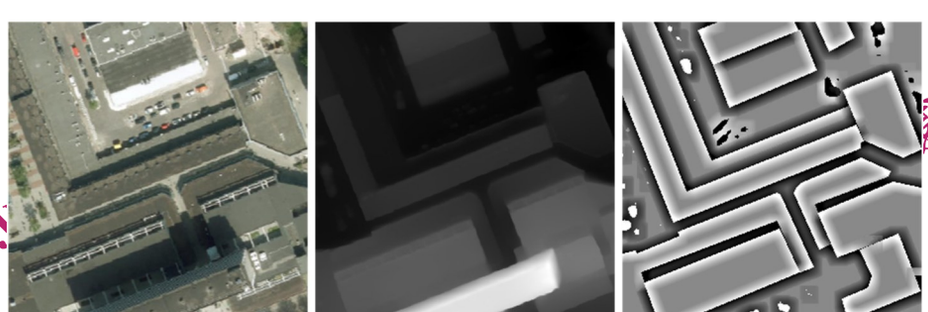
在复杂城市场景的高分辨率遥感影像中，高层建筑和桥梁等三维结构广泛存在。传统基于二维纹理的配准方法在此类场景下通常仅依赖平面信息，未利用地物的空间结构特征而导致配准精度受限。为了在城市多模态高分辨率遥感影像配准中有效利用三维空间信息，本文采用单目深度估计从影像中提取相对深度信息。然后基于所获深度在匹配过程中进行三维几何一致性约束，提高所提方法在复杂城市环境下的配准精度与鲁棒性。与激光雷达或多目相机等依赖额外传感器的深度获取方法相比，单目深度估计仅需单幅影像输入，具有数据易获取、适用性强等特点（Rajapaksha等，2024）。

Depth Anything（Yang等，2024）经过两个版本的迭代，已具有稳定的单目深度估计能力，并在数据生成（Hong等，2025；Ren等，2025）、新视角合成（Szymanowicz等，2025）、三维重建（Yu等，2025）和自动驾驶（Jia等，2024；Shi等，2025）等复杂场景中得到广泛应用，表现出良好的泛化能力和的鲁棒性（Chen等，2024）。需要注意的是，对于同一地物，其真实的物理高度是确定的。但由于影像获取时相、观测角度和成像条件的差异，基于单幅影像估计得到的深度值可能存在一定差别（见图5）。为降低这种差异对后续匹配的负面影响，本文使用Depth Anything V2分别对参考影像和待配准的跨模态翻译影像处理，生成对应的深度图。然后通过深度对齐策略获得统计分布一致的对齐深度图。具体而言，首先对原始深度图进行16×16邻域的局部标准化，该邻域尺寸在保持局部结构与统计稳定性之间取得了较好的平衡。随后采用2%百分位截断来抑制离群值的影响，最后利用经典的直方图匹配方法（Gonzalez和Woods，2018）对深度分布进行对齐，减弱不同影像间的深度偏移。该对齐深度图为后续特征匹配和几何约束提供了可靠的数据基础和空间先验。



（a） 待配准伪光学影像及其深度图与对齐深度图

（a） Pseudo image to be registered with its depth and aligned depth maps



（b） 参考光学影像及其深度图与对齐深度图

（b） Reference optical image with its depth and aligned depth maps

图5 获取对齐深度图

Fig. 5 Acquisition of the aligned depth map

2.3 基于深度的三维空间关系约束匹配方法

2.3.1 多特征图联合特征点检测

传统配准方法通常在原始灰度影像，或为应对模态差异在各向异性加权特征图上进行特征点检测。本文在2.1节中已将异源遥感影像转换为统一模态，在特征点检测时无需再考虑模态差异。然而，仅在单一灰度图或特征图上进行特征点检测，可能无法充分捕获影像的纹理、边缘和结构信息。为此，本文提出了一种基于多特征图联合的特征点检测方法，以获取城市高分辨率遥感影像中潜在的同名特征点。

具体而言，先对影像进行高斯平滑，再利用一阶微分算子（如Sobel或Prewitt）（何等，2018）计算影像在x和y方向上的梯度分量∇x和∇y，进而生成梯度幅度图和梯度方向图。将原始影像和这两个特征图构建为特征点候选空间。在此空间内，对每个特征图使用FAST算法进行特征点检测，以获取对纹理细节、边缘强度与方向信息敏感的特征点集合。该方法能够在城市复杂场景中更好地捕捉多维信息，为后续的稳健特征匹配提供丰富的特征点候选集。

2.3.2 深度联合特征描述符构建

构建具有鲁棒性的特征描述符是实现高精度配准的关键。本文借鉴经典的GLOH描述符（Mikolajczyk和Schmid，2003）和HAPCG描述符（姚等，2021）框架，提出了一种深度联合特征描述符，该描述符既包含影像的二维纹理特征又融合了三维空间特征，对高分辨城市复杂环境遥感影像配准具有较强的鲁棒性。

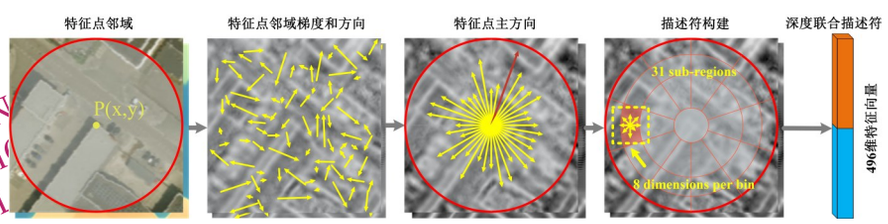


图6 深度联合特征描述符构建流程

Fig. 6 Construction process of the depth-joint feature descriptor

深度联合特征描述符的构建流程如图6所示：首先以特征点为中心选取固定大小的圆形描述符邻域，在该邻域内计算绝对相位一致性梯度和方向特征。随后将整个邻域按照10°为间隔等分为36份，统计每一份中的梯度特征和方向特征，并将统计的峰值方向作为该特征点的主方向。然后以特征点为极点、以主方向为极轴建立对数极坐标系，将整个描述符邻域划分为3×10+1=31个子区域。其中径向根据半径被划分为4个区域（一个中心圆和3个同心环），同心环按照36°为间隔等分为10个等角扇区。最后统计这31个子区域内每个像素点的8个方向的梯度和方向特征，生成大小为31×8=248维的纹理特征描述向量。与此同时，在对应的深度图上进行相同的描述符构建流程，生成248维深度特征描述向量。最后，将这两个特征描述向量串联并进行归一化，形成一个既包含纹理特征又包含三维信息的496维联合描述符。

2.3.3 基于深度的三维空间关系约束特征匹配

完成特征点提取与描述后，需要对特征点进行粗略匹配和精筛提高匹配精度。本文首先采用最近邻搜索策略，为待配准影像中的特征点在参考影像中寻找描述符距离最近的候选点，并通过坐标去重以确保一对一的匹配关系。然而，仅依赖描述符相似性的匹配通常会产生大量误匹配，即使是在引入深度联合特征描述符的情况下，该问题仍难以完全避免。这是因为部分特征点在二维影像或描述符空间中表现相似，但在三维空间中并非对应的同一地物点。这些误匹配点通常在对应的深度图中表现出明显的深度不一致。

为解决以上问题，本文提出了一种基于多尺度深度一致性约束的匹配点筛选策略。该策略利用对齐深度图提供的伪三维信息，对初始匹配结果进行几何一致性校验，它主要包括以下几个步骤：

（1）多尺度邻域深度计算。对于每一对候选匹配点对，分别在参考影像和待配准影像的对齐

深度图上选取不同尺寸大小（如3×3，5×5和7×7）的邻域窗口，随后计算两幅影像中对应邻域的平均深度绝对差值，取其最大值作为该匹配点对的多尺度深度差度量，并将其标记为：

$$\Delta_i = \max\left(\bar{D}_A^k - \bar{D}_B^k\right) \tag{2}$$

其中，$\bar{D}_A^k$和$\bar{D}_B^k$分别表示在尺度k下两幅深度图对应邻域的平均深度，Δ_i表示第i个候选匹配点对的最大深度偏差。

（2）自适应阈值计算。为了自适应地判别异常匹配点，本文采用基于中位数和中位绝对偏差的统计策略来确定阈值T：

$$\begin{cases} T = \text{median}(\boldsymbol{\Delta}) + n \cdot MAD(\boldsymbol{\Delta}) \\ MAD(\boldsymbol{\Delta}) = \text{median}\left(\left|\Delta_i - \text{median}(\boldsymbol{\Delta})\right|\right) \end{cases} \tag{3}$$

其中，$\boldsymbol{\Delta}$表示所有候选匹配点对的最大深度偏差集合，n为阈值系数。经测试发现，当$n\in[1, 4.5]$时，NCM随n的增大稳步增加，对应的RMSE_{all}值略微上升但整体波动较小。为兼顾匹配点数量和配准精度，本文取n=3，该值在统计意义上与鲁棒统计学中的3σ准则思想相近。

（3）深度一致性筛选和误差剔除。若某匹配点对的深度差度量$\Delta_i > T$，则认为该匹配存在显著几何不一致性，将其视为误匹配并剔除。最后，利用当前内点估计变换矩阵$\boldsymbol{H}$并计算待配准点的重投影误差，保留前25%的低偏差特征点，并利用$\boldsymbol{H}$修正其余点的坐标，随后迭代计算变换矩阵$\boldsymbol{H}$以实现亚像素级定位。通过以上步骤，可有效去除二维纹理特征相似，但因地形起伏、视角差或遮挡导致的错误匹配，提升配准的几何精度与结构一致性。

3 实验与结果分析

3.1 实验数据

为全面验证所提方法的有效性与鲁棒性，本文选取了4种不同组合的异源遥感影像作为测试数据，它涵盖了多种模态差异和城市形变类型（见图7）。这些高分辨率的测试影像采集自DDHRNet数据集、SpaceNet6竞赛数据集以及公开的地图服务平台（Esri World Imagery、Google Earth等）。测试数据覆盖了典型的城市复杂场景，包括密集建筑、高层建筑以及阴影区域，旨在最大程度地模拟真实城市环境下遥感影像配准所面临的挑战。根据影像形变机制的不同，这四组测试数据可划分为两类：人为扰动的几何形变数据与自然存在的形变数据：

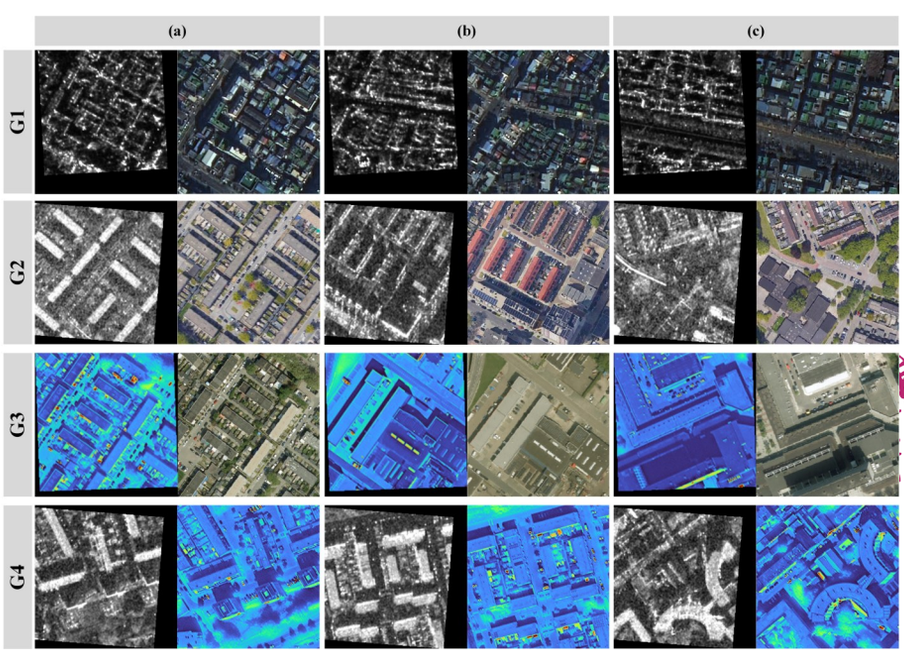


图7 本文的实验数据

Fig. 7 Experimental data used in this study

（1）人为扰动的几何形变数据。包括G1（SAR-RGB）和G4（SAR-IR）两组，其原始数据分别来自DDHRNet数据集（韩国浦项市）和SpaceNet 6竞赛数据集。由于上述数据集中的影像

在发布前已完成预配准，缺乏显著的几何差异，不足以反映算法在复杂形变条件下的表现。为此，本文在待配准影像上引入随机透视变换矩阵，人为施加平移、旋转和缩放等扰动，以模拟实际遥感配准任务中常见的几何形变。其中，旋转角度在［-2.5°，2.5°］范围内随机采样，缩放比例在［0.85，1.15］范围内变化，平移量在x和y方向均控制在［-10，10］像素范围内。所有参数均采用均匀分布随机生成。

（2）自然存在的形变数据。包括G2（SAR-RGB）和G3（IR-RGB）两组，它们采集自SpaceNet 6和公开地图服务（Esri World Imagery与Google Earth）。这些影像在发布前虽然经过预校正，但由于传感器类型、成像视角和采集时间的差异，影像对之间表现出明显的跨模态差异、视角视差及自然几何形变。这两组数据能够真实反映城市影像配准中面临的挑战。

通过对两类数据的实验对比，不仅能够验证所提方法对特定形变的纠正能力，也能评估其在真实复杂城市场景下的鲁棒性与泛化性能。各组测试数据的详细信息请参考表2。

表2　本文测试数据详表

Table 2　Detailed information of the test data used in this study

Group	Image Source	Image Provider	GSD/m	Image Size/pixel	Date
G1 (SAR-RGB)	GF3	DDHRNet dataset	1.0	256×256	---
	GF2	DDHRNet dataset	1.0	256×256	
G2 (SAR-RGB)	Capella Space	SpaceNet 6	0.5	256×256	2019.8.23
	--- (various)	Google Earth	0.45	300×300	2021.5.29/2022.5.8
G3 (IR-RGB)	WorldView-2	SpaceNet 6	0.5	236×236	2019.8.31
	--- (various)	Esri World Imagery	0.45	300×300	2024.5.12/2024.7.29
G4 (SAR-IR)	Capella Space	SpaceNet 6	0.5	256×256	2019.8.23
	WorldView-2	SpaceNet 6	0.5	256×256	2019.8.31

3.2　评价指标

为了客观、系统地评价所提方法的性能，本文从定性和定量两个方面对配准结果进行综合评估。本文使用棋盘格图（Karantzalos等，2014）来进行定性评价，通过观察方格边界是否连续来判断不同方法在细节处理上的优劣。由于本文对比的其他方法在编写语言和是否使用线程池上有所差异，因此本文主要通过正确匹配的数量（number of correct matches，NCM）、正确匹配率（Correct Match Ratio，CMR）以及均方根误差（Root Mean Square Error，RMSE）来定量评估几种方法的性能，它们的单位分别为“个”、“%”和“像素”（Xiang等，2018；Yan等，2022）。

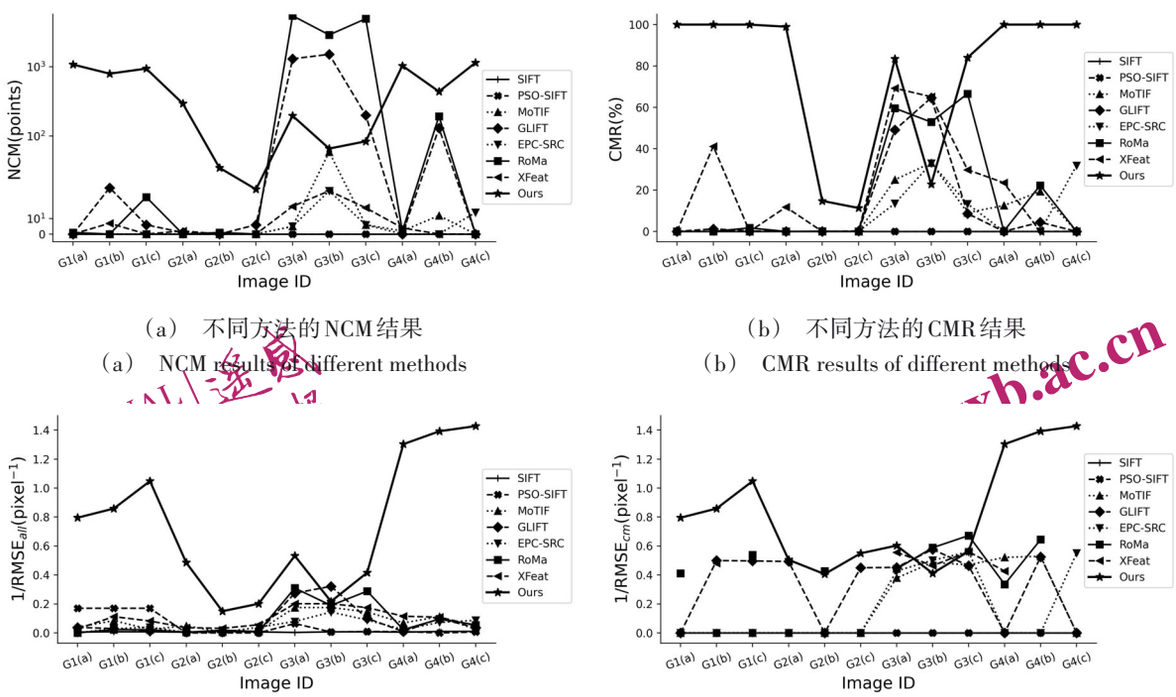
对于G1和G4实验组，保存人为扰动的变换矩阵作为地面真值；对于G2和G3实验组，人工采集超过15个良好的控制点，计算透视变换矩阵作为地面真值，并将距真值3个像素内的匹配点标记为正确匹配点。$RMSE_{all}$为配准方法输出的所有配准点对计算得到的RMSE，$RMSE_{cm}$为仅使用正确匹配点对计算得到的RMSE。

3.3　对比实验

本文将所提方法与EPC-SRC（郑等，2025）、GLIFT（Zhang等，2025）、MoTIF（Yao等，2022）、PSO-SIFT（Ma等，2017）、SIFT（Lowe，2004）等传统方法及XFeat（Potje等，2024）、RoMa（Edstedt等，2024）等深度学习方法进行对比测试，以评估所提方法的配准性能。所有测试代码收集自开源网络，并保持默认的最优参数。本测试是在一台小米笔记本电脑上进行的，它装备了英特尔（R）酷睿（TM）i7-9750H CPU，16GB RAM和Windows 10 x64操作系统。本文方法已在MATLAB R2023b上通过了测试。

图8为本文方法与其他7种方法的定量结果。为便于直观比较，图8（c）和（d）分别取$RMSE_{all}$和$RMSE_{cm}$的倒数，此时低值代表配准精度差，高值代表精度好。综合NCM和$RMSE_{all}$来看，除G3（b）测试项外，所提方法的配准精度均优于其他方法。除SIFT和PSO-SIFT外的其他方法次之，但也仅在G3测试组中的匹配效果较好，这是因为在配准算法中，需要先将影像转化为灰度再进行配

准，而G3测试组中红外和光学影像的灰度图极为相似，可以看作是同模态影像匹配。SIFT和PSO-SIFT方法的配准效果最差，几乎没有正确匹配点对。此外，GLIFT和RoMa在G3测试组中虽取得了可观的NCM，但配准精度较本文方法并无显著提升。上述结果表明，所提方法在复杂场景下具有更强的鲁棒性，在提高匹配数量的同时有效控制了匹配误差，实现了两者之间的良好平衡。总体而言，所提方法的平均NCM为510，较其他方法提升约0.92~5.22倍；平均RMSE$_{all}$为1.58像素，精度提升约4.12~4.78倍。需要指出的是，所提方法的平均运行时间约为37秒，其中约25~35秒耗时于联合描述符构建。这是因为当前实现为未经工程优化的算法原型，主要用于验证算法的设计和可行性，后续将进一步优化算法的效率。



（a） 不同方法的NCM结果

（a） NCM results of different methods

（b） 不同方法的CMR结果

（b） CMR results of different methods

（c） 不同方法的1/RMSE$_{all}$结果

（c） 1/*RMSE*$_{all}$ results of different methods

（d） 不同方法的1/RMSE$_{cm}$结果

（d） 1/*RMSE*$_{cm}$ results of different methods

图8 各方法在不同指标下的配准性能对比

Fig. 8 Comparison of registration performance metrics among different methods

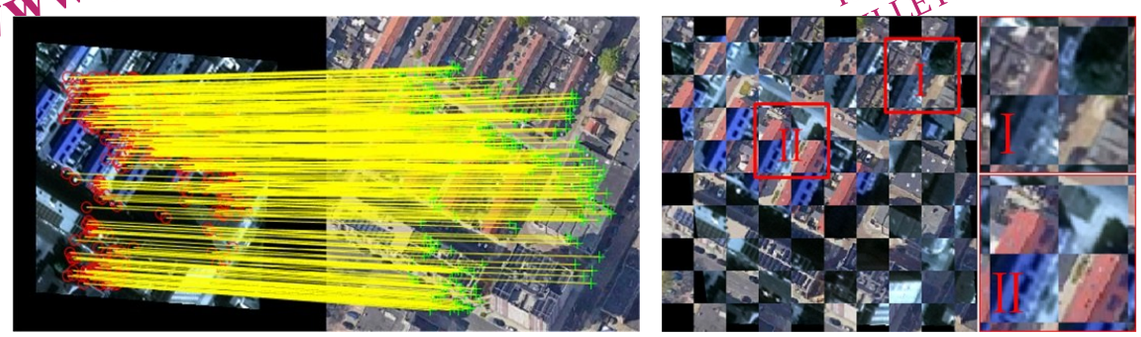
尽管所提方法的配准精度优于其他方法，但它在G2（b）、G2（c）和G3（b）测试项的正确匹配率较低，RMSE也出现了一定程度的下降。事实上，所提方法的棋盘格图在影像边缘也有较好的配准效果（见9（b）），这是因为现有的定量评估方法在复杂城市场景中存在局限。其根本原因在于，基于全局单应性矩阵的精度验证高度依赖于场景的平面假设。在这种情况下，即使在高层建筑、地形起伏及显著视差的复杂区域存在正确匹配点，也会因为偏离全局变换模型而被误判为离群点。因此，定量指标可能会低估算法在复杂三维环境下的真实配准能力。为弥补这一评估偏差，本文对G2（b）、G2（c）和G3（b）测试项引入了定性评估以更全面地验证配准效果。

通过对比图10中不同方法的棋盘格图可以看出，在G2（b）、G2（c）和G3（b）等典型城市场景下，本文方法具有更好的视觉连续性。这表明本文算法在处理具有显著高差和非线性形变影像时具有较强的稳健性。同时，该结果也反映出当前基于全局单应性矩阵评的价体系在复杂场景中的局限性，凸显出构建能够真实反映复杂地物结构的立体评价标准具有必要性。

3.4 消融实验

为验证本文所提方法中各关键步骤对配准性能的贡献，我们设计并实施了渐进式消融实验。

实验方案如下：方案1（Base）：基准方案，使用原始异源影像和GLOH描述符（Mikolajczyk和Schmid，2003）进行配准；方案2（Base+A）：在方案1的基础上引入模态转换；方案3（Base+A+B）：将方案2中的描述符替换为深度联合描述符；方案4（Base+A+B+C）：在方案3的基础上进行多尺度深度一致性约束，即本文所提方法。



（a） “G2(b)”所提方法匹配结果

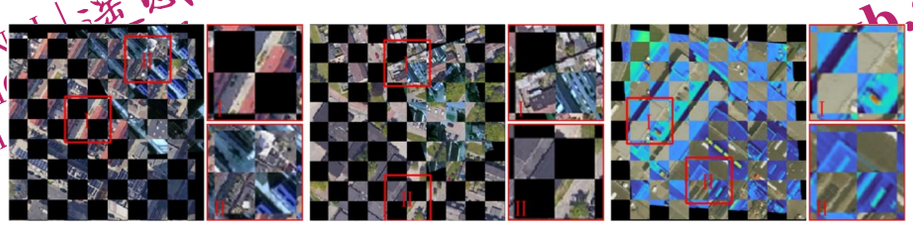
（a） Matching results of the proposed method for “G2(b)”

（b） 所提方法棋盘格图

（b） Checkerboard image generated by the proposed method

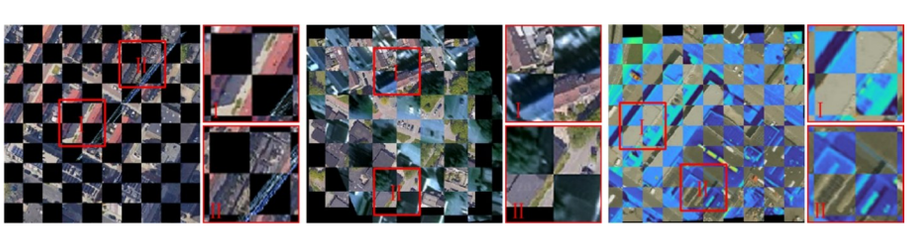
图9 基于全局变换矩阵的定量评价缺陷

Fig. 9 Deficiencies of quantitative evaluation based on the global transformation matrix



（a） MoTIF方法定性评价结果

（a） Qualitative evaluation results of the MoTIF method



（b） EPC-SRC方法定性评价结果

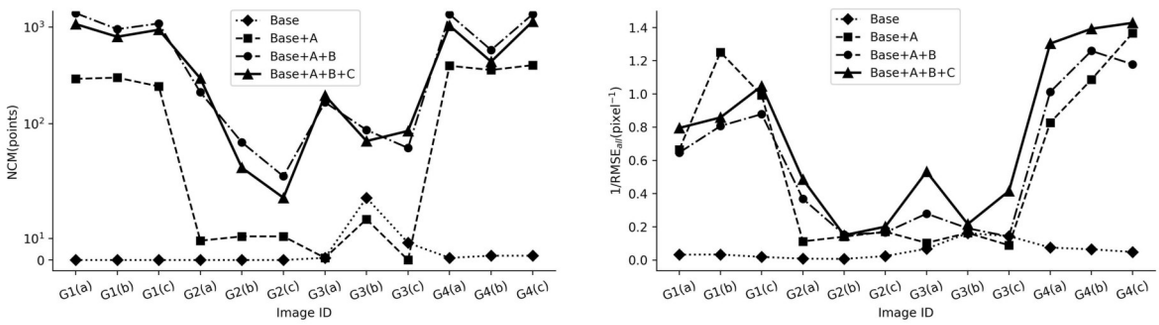
（b） Qualitative evaluation results of the EPC-SRC method

（c） 本文方法定性评价结果

（c） Qualitative evaluation results of the proposed method

图10 不同方法在相同测试项的定性评价结果

Fig. 10 Qualitative evaluation results of different methods on the same test case



（a） 匹配数量（NCM）消融结果　　（b） 配准精度（1/RMSE$_{all}$）消融结果

（a） Ablation results for matching quantity（NCM）　　（b） Ablation results for registration accuracy（1/RMSE$_{all}$）

图 11 消融实验结果

Fig. 11 Results of ablation experiments

消融实验的结果如图11所示，实验分析表明：模态转换通过统一影像的表征空间，显著提升了正确匹配点数（NCM）并降低了RMSE$_{all}$，但在自然形变数据组上的提升有限。深度联合描述符的特征描述能力更强，提升了算法在自然形变数据组上的配准性能。多尺度深度一致性约束通过剔除潜在的误匹配点，在保证NCM的前提下，进一步提高了算法的配准的精度，实现匹配点数量与精度的平衡。综上，本文方法通过整合以上关键步骤，在相同的测试场景下具有更优的综合性能。

3.5 鲁棒性实验

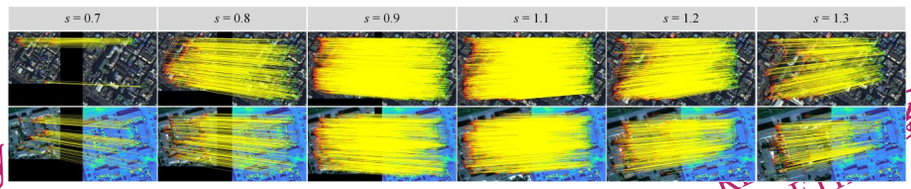
为验证所提方法在尺度、旋转及其组合扰动下的鲁棒性，分别在SAR-光学和SAR-红外测试组中分别选取一对影像对进行实验。

（1）尺度鲁棒性测试。以0.1为间隔对生成的伪光学影像进行缩放，生成尺度范围为0.7~1.3的6组模拟图像。匹配结果如图12（a）所示，随着尺度差异的增大，匹配点数量呈逐渐下降的趋势。在整个测试范围内，所提方法能够获得至少80对匹配点和1.6个像素内的匹配误差。

（2）旋转鲁棒性测试。以10°为间隔进对为伪光学影像进行顺时针和逆时针旋转，获得旋转范围为-30°到30°的6组模拟图像。匹配结果如图12（b）所示，角度在-20°至20°范围内时，方法仍可获得超过150对有效匹配点，RMSE约为1.9像素。当旋转角度进一步增大时，匹配性能逐渐下降。

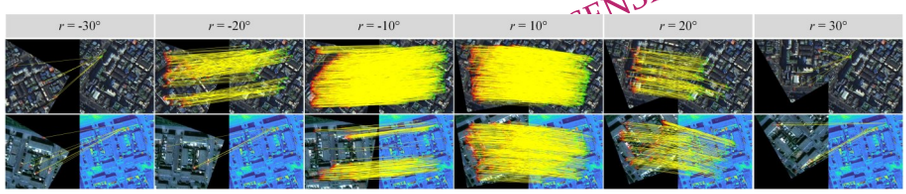
（3）组合扰动测试。在以上基础上，同时引入尺度、旋转和平移扰动（平移步长为7像素，范围为-21至21像素），构建综合变换测试数据。匹配结果如图12（c）所示，当尺度变化在0.9到1.2、旋转角度在-10°至20°、平移范围在-7至14像素内，所提方法仍可获得稳定的匹配结果。在该区间内，匹配点数量不少于19对，RMSE控制在0.45－2.26像素之间。

综上，所提方法在中等强度的尺度、旋转以及组合扰动下均表现出良好的鲁棒性，能够适应实际遥感场景中常见的成像尺度变化和视角差异。但在较大幅度复合变换的条件下匹配性能会有所下降，表明所提方法在城市复杂形变下仍存在一定局限性，有待进一步优化。



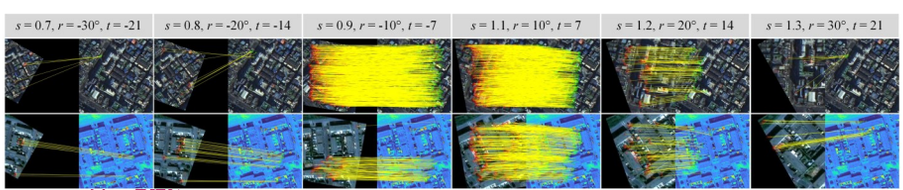
（a） 尺度鲁棒性测试结果

（a） Evaluation of scale robustness



（b） 旋转鲁棒性测试结果

（b） Evaluation of rotational robustness



（c） 组合扰动测试结果

（c） Results of composite perturbation evaluation

图 12 鲁棒性测试结果（s：尺度；r：角度；t：平移）

Fig. 12 Robustness test results (s: scale; r: rotation; t: translation)

4 结 论

针对复杂城市环境下多模态高分辨率遥感影像配准难度增大、三维空间信息利用不足等问题，本文提出了一种基于模态转换和三维空间关系约束的多模态遥感影像配准方法。在包含人为扰动和自然形变的4组城市影像数据上，与EPC-SRC、GLIFT、MoTIF、PSO-SIFT、SIFT等传统方法及XFeat、RoM等深度学习方法进行了对比实验。结果表明，所提方法的平均NCM为510，较其他方法提升约0.92~5.22倍；RMSE平均为1.58像素，精度提升约4.12~4.78倍。这表明所提方法在复杂场景下具有更强的鲁棒性，在提高匹配数量的同时有效控制了匹配误差，实现了两者之间的良好平衡。

尽管如此，本文方法的平均运行时间相对较长，这是因为当前的代码实现为未经工程优化的算法原型，旨在验证算法的可行性。在下一步工作中，我们将通过代码优化、GPU加速和并行计算等手段提升算法的运行效率。此外，实验分析表明，现有基于全局单应性矩阵的精度评价体系在评估高差显著或视差明显的复杂区域时存在一定局限性。因此，有必要构建更加契合城市复杂空间结构特征的综合评价框架。进一步研究还可探索三维重建与配准的联合优化的策略，实现从二维像素空间到真实三维几何空间的协同建模，为城市多源遥感影像的智能融合和空间认知提供更坚实的基础。

参考文献（References）

Chen J M, Liu B, Yu A Z, Quan Y J, Li T T and Guo W Y. 2024. Depth feature fusion network for building extraction in remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 16577-16591 [DOI: 10.1109/JSTARS.2024.3452640]

Edstedt J, Sun Q Y, Bökman G, Wadenbäck M, Felsberg M. 2024. RoMa: Robust dense feature matching. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 19790-19800 [DOI: 10.1109/CVPR52733.2024.01871]

Gonzalez R C and Woods R E. 2018. Digital Image Processing. New York: Pearson.

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D,

Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks. arXiv [DOI: 10.48550/arXiv.1406.2661].

He M M, Guo Q, Li A, Chen J, Chen B and Feng X X. 2018. Automatic fast feature-level image registration for high-resolution remote sensing images. National Remote Sensing Bulletin, 22(2): 277-292 (何敏章, 郭擎, 李安, 陈俊, 陈勃, 冯旭祥. 2018. 特征级高分辨率遥感图像快速自动配准. 遥感学报, 22(2): 277-292 [DOI: 10.11834/jrs.20186420])

Hong Z C, Wu T, Xu Z Y and Zhao W F. 2025. Depth2Elevation: scale modulation with Depth Anything model for single-view remote sensing image height estimation. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-14 [DOI: 10.1109/TGRS.2025.3564820].

Hou Z L, Liu Y X and Zhang L. 2024. POS-GIFT: A geometric and intensity-invariant feature transformation for multimodal images. Information Fusion, 102: 102027 [DOI: 10.1016/j.inffus.2023.102027].

Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 5967-5976 [DOI: 10.1109/CVPR.2017.632].

Karantzalos K, Sotiras A and Paragios N. 2014. Efficient and automated multimodal satellite data registration through MRFs and linear programming. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, OH, USA: IEEE: 335-342 [DOI: 10.1109/CVPRW.2014.57].

Kim H and Kim M G. 2012. Image registration using terrain relief correction based on rigorous sensor models. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX-B1: 235-238 [DOI: 10.5194/isprsarchives-XXXIX-B1-235-2012].

Li J Y, Hu Q W and Ai M Y. 2020. RIFT: multi-modal image matching based on radiation-variation insensitive feature transform. IEEE Transactions on Image Processing, 29: 3296-3310 [DOI: 10.1109/TIP.2019.2959244].

Li Y L, Mo Z B and Di X. 2024. SafeAug: safety-critical driving data augmentation from naturalistic datasets. 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). 3251-3256 [DOI: 10.1109/ITSC58415.2024.10920171].

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94].

Ma W P, Wen Z L, Wu Y, Jiao L C, Gong M G, Zheng Y F and Liu L. 2017. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geoscience and Remote Sensing Letters, 14(1): 3-7 [DOI: 10.1109/LGRS.2016.2600858].

Mikolajczyk K and Schmid C. 2003. A performance evaluation of local descriptors. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.: Vol. 2. Madison, WI, USA: IEEE Comput. Soc: II-257-II-263 [DOI: 10.1109/CVPR.2003.1211478].

Potje G, Cadar F, Araujo A, Martins R, Nascimento E R. 2024. XFeat: Accelerated features for lightweight image matching. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 2682-2691 [DOI: 10.1109/CVPR52733.2024.00259].

Rajapaksha U, Sohel F, Laga H, Diepeveen D and Bennamoun M. 2024. Deep learning-based depth estimation methods from monocular image and videos: a comprehensive survey. ACM Comput. Surv., 56(12): 315:1-315:51 [DOI: 10.1145/3677327].

Ren B, Ma S B, Hou B, Hong D F, Chanussot J, Wang J L and Jiao L C. 2022. A dual-stream high resolution network: deep fusion of GF-2 and GF-3 data for land cover classification. International Journal of Applied Earth Observation and Geoinformation, 112: 102896 [DOI: 10.1016/j.jag.2022.102896].

Ren J W, Jiang X Y, Li Z Z, Liang D K, Zhou X and Bai X. 2025. Modality invariant image matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 23059-23068.

Rosten E, Porter R and Drummond T. 2010. Faster and better: a machine learning approach to corner detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1): 105-119 [DOI: 10.1109/TPAMI.2008.275].

Sara U, Akter M and Uddin M S. 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. Journal of Computer and Communications, 07(03): 8-18 [DOI: 10.4236/jcc.2019.73002].

Shermeyer J, Hogan D, Brown J, Van Etten A, Weir N, Pacifici F, Hansch R, Bastidas A, Soenen S, Bacastow T and Lewis R. 2020. SpaceNet 6: multi-sensor all weather mapping dataset. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, WA, USA: IEEE: 768-777 [DOI: 10.1109/CVPRW50498.2020.00106].

Shi D H, Zhao C X, Zhao C B, Fang Z, Yu C H, Li J and Feng M J. 2025. Depth-aware unpaired image-to-image translation for autonomous driving test scenario generation using a dual-branch GAN. Frontiers in Neurorobotics, 19 [DOI: 10.3389/fnbot.2025.1603964].

Sun Y, Wang Y, Eineder M. 2024. QuickQuakeBuildings: Post-Earthquake SAR-optical dataset for quick damaged-building detection. IEEE Geoscience and Remote Sensing Letters, 21: 1-5 [DOI: 10.1109/LGRS.2024.3406966]

Szymanowicz S, Insafutdinov E, Zheng C, Campbell D, Henriques J F, Rupprecht C and Vedaldi A. 2025. Flash3D: feed-forward generalisable 3D scene reconstruction from a single image. 2025 International Conference on 3D Vision (3DV). 670-681 [DOI: 10.1109/3DV66043.2025.00067].

Wang M M, Ye Y X, Zhu B and Zhang G. 2022. An automatic registration method for optical and SAR images based on spatial constraint and structure features. Geomatics and Information Science of Wuhan University, 47(1): 141-148 (王蒙蒙, 叶沅鑫, 朱柏, 张过. 2022. 基于空间约束和结构特征的光学与SAR影像配准. 武汉大学学报(信息科学版), 47(1): 141-148 [DOI: 10.13203/j.whugis20190354])

Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J and Catanzaro B. 2018. High-resolution image synthesis and semantic manipulation with conditional GANs. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 8798-8807 [DOI: 10.1109/CVPR.2018.00917].

Wells W M, Viola P, Atsumi H, Nakajima S and Kikinis R. 1996. Multi-modal volume registration by maximization of mutual information. Medical Image Analysis, 1(1): 35-51 [DOI: 10.1016/S1361-8415(01)80004-9].

Wu P H, Yao Y X, Zhang W F, Wei D, Wan Y, Li Y S and Zhang Y J. 2025. MapGlue: multimodal remote sensing image matching. arXiv [DOI: 10.48550/arXiv.2503.16185].

Xia W H, Zhang Y L, Yang Y J, Xue J H, Zhou B L and Yang M X. 2023. GAN inversion: asurvey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3121-3138 [DOI: 10.1109/TPAMI.2022.3181070].

Xiang Y M, Wang F and You H J. 2018. OS-SIFT: a robust SIFT-like algorithm for high-resolution optical-to-SAR image registration in suburban areas. IEEE Transactions on Geoscience and Remote Sensing, 56(6): 3078-3090 [DOI: 10.1109/TGRS.2018.2790483].

Xiong X, Jin G W, Cui R B, Xu M and Yang H. 2025. Optical-to-SARimage matching based on fast rank-based local self-similarity descriptor. 2025 6th International Conference on Geology, Mapping and Remote Sensing (ICGMRS). 545-549 [DOI: 10.1109/ICGMRS66001.2025.11065913].

Yan H, Yang S W, Xue Q and Zhang N X. 2022. HR optical and SAR image registration using uniform optimized feature and extend phase congruency. International Journal of Remote Sensing, 43(1): 52-74 [DOI: 10.1080/01431161.2021.1999527].

Yang C, Liu C, Tang T F and Ye Y X. 2025. Robust matching of optical and SAR images based on deep feature reconstruction enhancement. National Remote Sensing Bulletin, 29(8): 2616-2626 (杨超, 刘畅, 唐腾峰, 叶沅鑫.2025.基于深度特征重构增强的光学和SAR图像鲁棒匹配. 遥感学报, 29(8): 2616-2626 [DOI: 10.11834/jrs.20254295])

Yang L H, Kang B Y, Huang Z L, Xu X G, Feng J S and Zhao H S. 2024. Depth Anything: unleashing the power of large-scale unlabeled data. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE: 10371-10381 [DOI: 10.1109/CVPR52733.2024.00987].

Yao Y X, Zhang Y J, Wan Y, Liu X Y and Guo H Y. 2021. Heterologous images matching considering anisotropic weighted moment and absolute phase orientation. Geomatics and Information Science of Wuhan University, 46(11): 1727-1736 (姚永祥, 张永军, 万一, 刘欣怡, 郭浩宇. 2021. 顾及各向异性加权力矩与绝对相位方向的异源影像匹配. 武汉大学学报(信息科学版), 46(11): 1727-1736 [DOI: 10.13203/j.whugis20200702])

Yao Y, Zhang B, Wan Y and Zhang Y. 2022. Motif: multi-orientation tensor index feature descriptor for SAR-optical image registration. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2022: 99-105 [DOI: 10.5194/isprs-archives-XLIII-B2-2022-99-2022].

Ye Y B, Wang Q W, Zhao H, Teng X C, Bian Y J and Li Z. 2024. Fast and robust optical-to-SAR remote sensing image registration using region-aware phase descriptor. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-12 [DOI: 10.1109/TGRS.2024.3379370].

Ye Y X, Bruzzone L, Shan J, Bovolo F and Zhu Q. 2019. Fast and robust matching for multimodal remote sensing image registration. IEEE Transactions on Geoscience and Remote Sensing, 57(11): 9059-9070 [DOI: 10.1109/TGRS.2019.2924684].

Ye Y X and Shen L. 2016. HOPC: a novel similarity metric based on geometric structural properties for multi-modal remote sensing image matching. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, III-1: 9-16 [DOI: 10.5194/isprsannals-III-1-9-2016].

Ye Y X, Wang M M, Yang C, Yu Z R and Ge X M. 2024. Multisensor remote sensing registration method and system based on dense feature of orientated phase. National Remote Sensing Bulletin, 28(6): 1525-1538 (叶沅鑫, 王蒙蒙, 杨超, 喻智睿, 葛旭明. 2024. 基于方向相位稠密特征的多传感器遥感影像配准方法和系统. 遥感学报, 28(6): 1525-1538 [DOI: 10.11834/jrs.20221765])

Yu A Z, Hong D Y, Lu X B, Ji S and Fan J Y. 2025. Learning-based multiview stereo for remote sensed imagery with relative depth. IEEE Geoscience and Remote Sensing Letters, 22: 1-5 [DOI: 10.1109/LGRS.2025.3527560].

Yu J, Du S S, Xie G C, Lu R J, Li P W, Cai Z P and Lu K D. 2024. SAR2EO: a high-resolution image translation framework with denoising enhancement. AI 2023: Advances in Artificial Intelligence. Singapore: Springer Nature: 91-102 [DOI: 10.1007/978-981-99-8388-9_8].

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068].

Zhang S C, Luo B, Liu J, Fu Z T, Su X and Zhu S L. 2025. GLIFT: A global-to-local invariant feature transformation method for multimodal remote sensing image matching. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-19 [DOI: 10.1109/TGRS.2025.3599445].

Zhang Y J, Wu P H, Yao Y X, Wan Y, Zhang W F, Li Y S and Yan X. 2025. Multimodal remote sensing image robust matching based on second-order tensor orientation feature transformation. IEEE Transactions on Geoscience and Remote Sensing, 63: 1-14 [DOI: 10.1109/TGRS.2025.3535154].

Zheng H Q M, Pan C Y, Jin X, Wang Q Q, Miao S F and Jiang Q. 2025. Research progress of generative adversarial networks in remote sensing image fusion. National Remote Sensing Bulletin, 29(10): 2891-2904 (郑黄齐眉, 潘成毅, 金鑫, 王倩倩, 苗圣法, 江倩. 2025. 生成对抗网络在遥感图像融合中的研究进展. 遥感学报, 29(10): 2891-2904 [DOI: 10.11834/jrs.20254439])

Zheng Y, Yang S W, Li Y K, Wu J S, Shi Z and Kou R X. 2024. A heterogeneous remote sensing image matching method for urban ar-

eas with complex terrain based on 3D spatial relationship constraints. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17: 6791-6804 [DOI: 10.1109/JSTARS.2024.3374327]

Zheng Y, Yang S W, Wu J S, Fu Y kai and Kou R X. 2025. Heterogeneous remote sensing image registration based on extended phase consistency feature and spatial relationship constraint. Remote Sensing Technology and Application, 40(1): 144-155 (郑耀, 杨树文, 武锦沙, 付昱凯, 寇瑞雄. 2025. 基于扩展相位一致性特征和空间关系约束的异源遥感影像配准方法. 遥感技术与应用, 40(1): 144-155 [DOI: 10.11873/j.issn.1004-0323.2025.1.0144])

Zhou R F, Quan D, Wang S, Lv C H, Cao X W, Chanussot J, Li Y, Jiao L C. 2024. A unified deep learning network for remote sensing image registration and change detection. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-16 [DOI: 10.1109/TGRS.2023.3344751]

Multimodal High-Resolution Remote Sensing Image Registration via Modality Translation and 3D Spatial Relationship Constraints

LI Xiaokui[1], YANG Shuwen[1,2,3], LI Ziyuan[1], WANG Wenju[1], ZHU Hao[1], XUE Yiming[1]

1.Faculty of Geomatic, Lanzhou Jiaotong University, Lanzhou730070, China;
2.National-Local Joint Engineering Research Center of Technologies and Applications for National Geographic State Monitoring, Lanzhou 730070, China;
3.Gansu Province Key Laboratory of Science and Technology in Surveying & Mapping, Lanzhou 730070, China

Abstract: In complex urban environments, dense buildings and significant height variations amplify nonlinear radiometric differences and geometric distortions among multimodal remote sensing images. These challenges severely hinder high-precision registration of high-resolution imagery. Existing registration approaches predominantly rely on two-dimensional features such as texture, often neglecting the valuable three-dimensional spatial information inherent in such scenes. To address these limitations, this study proposes a novel multimodal high-resolution remote sensing image registration method that integrates modality transformation with three-dimensional spatial relationship constraints, aiming to improve robustness and accuracy in complex urban scenarios. The proposed method consists of three main components. First, a cross-modal image translation technique is employed to reduce radiometric discrepancies between multimodal images, effectively narrowing the modality gap and facilitating subsequent feature extraction. Second, monocular depth estimation is introduced to efficiently generate depth maps from single images. These depth maps provide essential spatial priors and serve as a foundation for constructing more informative feature descriptors and enforcing spatial constraints. Finally, a matching strategy based on depth-guided 3D spatial relationship constraints is developed. This strategy includes multi-feature map keypoint detection to capture potential salient features, the construction of depth-enhanced joint descriptors to improve feature distinctiveness, and the incorporation of 3D spatial relationship constraints to ensure geometrically consistent matching. Together, these steps enable reliable detection, robust description, and accurate matching of feature points across multimodal images. The proposed method was compared with several traditional and deep learning methods on four representative multimodal remote sensing datasets. Experimental results show that the proposed method achieves an average Number of Correct Matches (NCM) of 510, which improves upon existing methods by a factor of 0.92 to 5.22. The average Root Mean Square Error (RMSE) is 1.58 pixels, and the registration accuracy is improved by a factor of 4.12 to 4.78 compared to state-of-the-art methods. These results demonstrate that the proposed method has clear advantages in registration accuracy and overall performance, effectively suppressing nonlinear radiometric differences and geometric distortions in urban multimodal images and achieving high-precision registration. This study presents a robust solution for multimodal high-resolution remote sensing image registration in complex urban environments. By combining cross-modal translation, monocular depth estimation, and depth-based three-dimensional spatial relationship constraints, the proposed method successfully addresses both nonlinear radiometric differences and geometric distortions. The integration of 3D spatial information significantly improves feature matching robustness and registration precision compared to conventional 2D-based approaches. Experimental validation confirms that the method achieves superior performance in terms of both accuracy and stability, establishing a solid foundation for downstream applications such as urban mapping, change detection, and multi-sensor data fusion.

Key words: image registration, multimodal imagery, 3D spatial relationships, complex urban scenes, depth maps

Supported by Supported by National Natural Science Foundation of China (No. 42471438)