自动化学报
ACTA AUTOMATICA SINICA

# 基于运动过滤和调整的离群点移除

赖桃桃 张一凡 李佐勇 肖国宝 林维斯 王菡子

# Outlier Removal Based on Motion Filtering and Adjustment

LAI Tao-Tao, ZHANG Yi-Fan, LI Zuo-Yong, XIAO Guo-Bao, LIN Wei-Si, WANG Han-Zi

在线阅读 View online: https://doi.org/10.16383/j.aas.c250235

---

# 基于运动过滤和调整的离群点移除

赖桃桃[1] 张一凡[2] 李佐勇[1] 肖国宝[3] 林维斯[4] 王菡子[5]

**摘 要** 由现有的特征提取器建立的图像特征点匹配集合通常包含大量离群点, 这严重影响特征匹配的有效性和依赖匹配结果的下游任务的性能. 最近提出的几种离群点去除方法通过估计运动场来利用匹配对的运动一致性, 并使用卷积神经网络 (CNN) 来减少离群点造成的污染, 以捕获上下文. 然而, CNN 在捕捉全局上下文方面存在固有缺陷, 其感受野的固定性与局部性导致模型难以自适应地整合远距离信息, 从而制约相关方法的性能. 与这些使用卷积神经网络直接估计运动场的方法不同, 本文尝试在不使用 CNN 的情况下估计高质量的运动场. 因此, 提出基于运动过滤和调整的网络, 以减轻在捕捉上下文时离群点的影响. 具体而言, 首先, 设计一个运动过滤模块, 以迭代地去除离群点并捕获上下文. 然后, 设计一个规则化和调整模块, 该模块先估计初始运动场, 接着通过利用额外的位置信息对其进行调整, 使其更加准确. 在离群点去除和相对姿态估计任务中, 利用室内和室外数据集评估所提出方法的性能. 实验结果表明, 与现有多种方法相比, 所提方法展现出更优的性能.

**关键词** 计算机视觉; 离群点移除; 运动过滤; 规则化; 调整

**引用格式** 赖桃桃, 张一凡, 李佐勇, 肖国宝, 林维斯, 王菡子. 基于运动过滤和调整的离群点移除. 自动化学报, 2026, **52**(3): 593−610

**DOI** 10.16383/j.aas.c250235 **CSTR** 32138.14.j.aas.c250235

## Outlier Removal Based on Motion Filtering and Adjustment

LAI Tao-Tao[1] ZHANG Yi-Fan[2] LI Zuo-Yong[1] XIAO Guo-Bao[3] LIN Wei-Si[4] WANG Han-Zi[5]

**Abstract** The image point correspondences established by off-the-shelf feature extractors usually contain a large number of outliers, which severely affects the effectiveness of feature matching and the performance of downstream tasks reliant on the matching results. Several recently proposed outlier removal methods leverage the motion consistency of correspondences by estimating a motion field and employ convolutional neural network (CNN) to reduce contamination from outliers to capture context. However, CNN inherently suffers from limitations in capturing global context, as the fixed and localized nature of their receptive fields makes it difficult for models to adaptively integrate long-range information, thereby constraining the performance of related methods. Departing from these methods that directly estimate motion fields using CNN, this paper explores estimating a high-quality motion field without reliance on CNN. To this end, a motion filtering and adjustment network (MFANet) is proposed to mitigate the impact of outliers during context capture. Specifically, a motion-filtering block is first designed to iteratively remove outliers and capture contextual information. Then, a regularization and adjustment block is designed to estimate an initial motion field, which is then refined for greater accuracy by incorporating additional positional information. The performance of MFANet is evaluated on both indoor and outdoor datasets for the tasks of outlier removal and relative pose estimation. Experimental results demonstrate that MFANet achieves superior performance compared to several existing methods.

**Keywords** computer vision; outlier removal; motion filtering; regularization; adjustment

**Citation** Lai Tao-Tao, Zhang Yi-Fan, Li Zuo-Yong, Xiao Guo-Bao, Lin Wei-Si, Wang Han-Zi. Outlier removal based on motion filtering and adjustment. *Acta Automatica Sinica*, 2026, **52**(3): 593−610

收稿日期 2025-05-23 录用日期 2025-09-29

Manuscript received May 23, 2025; accepted September 29, 2025

国家自然科学基金 (62172197, 62471207, 62472312, U21A20514), 福建省自然科学基金 (2024J01209, 2024J02029), 福建省发树慈善基金会资助研究专项 (MFK24003) 资助

Supported by National Natural Science Foundation of China (62172197, 62471207, 62472312, U21A20514), Natural Science Foundation of Fujian Province (2024J01209, 2024J02029), and Research Project of Fashu Foundation (MFK24003)

本文责任编委 林宙辰

Recommended by Associate Editor LIN Zhou-Chen

1. 闽江学院计算机与大数据学院, 福建省信息处理与智能控制重点实验室 福州 350108 中国 2. 福州大学计算机与大数据学院 福州 350108 中国 3. 同济大学计算机科学与技术学院 上海 201804 中国 4. 南洋理工大学计算机科学与工程学院 新加坡 639798 新加坡 5. 厦门大学信息学院 厦门 361005 中国

1. Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, School of Computer and Data Science, Minjiang University, Fuzhou 350108, China 2. College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China 3. School of Computer Science and Technology, Tongji University, Shanghai 201804, China 4. School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore 5. School of Informatics, Xiamen University, Xiamen 361005, China

特征匹配旨在建立来自同一场景或相似场景的图像对的可靠特征点对应关系, 这是模式识别与图像处理中许多任务的基础, 如同时定位与地图构建[1]、消失点估计[2]、点云配准[3]、图像拼接[4]、运动分割[5]. 给定一对图像, 典型的特征匹配流程包括特征提取、特征匹配和离群点去除. 具体而言, 首先使用现有的特征提取器 (例如尺度不变特征变换 (scale-invariant feature transform, SIFT)[6] 和平方根归一化尺度不变特征变换 (root scale-invariant feature transform, RootSIFT)[7]) 从图像中获取特征点和描述符; 然后特征匹配基于对应描述符的相似性建立初始的对应关系 (匹配对); 最后离群点去除旨在识别和去除错误的匹配对 (即离群点), 同时尽可能多地保留正确的匹配对 (即内点).

许多复杂的场景增加了特征匹配任务的难度. 首先, 图像对可能包含重复结构、失真和遮挡. 其次, 图像对可能在视角、光照和深度方面有显著变化. 这些复杂的场景严重影响了初始匹配对的质量, 导致大量离群点的出现[8], 这给离群点去除带来挑战. 此外, 检测到的特征点通常聚集在纹理丰富的区域, 因此初始匹配对的分布通常是不均匀的, 并且会随着不同的图像场景而变化. 这使得从初始匹配对中收集可靠的信息变得困难.

在过去的几十年里, 传统方法为应对这些挑战做出重大努力. 传统方法包括以随机采样一致性方法 (random sample consensus, RANSAC)[9] 为代表的随机采样方法及其变体[10−12] 以及许多其他方法, 如局部性保持匹配方法 (locality preserving matching, LPM)[13]、向量场一致性方法 (vector field consensus, VFC)[14]、线性自适应滤波方法 (linear adaptive filtering, LAF)[15] 和基于网格的运动统计方法 (grid-based motion statistics, GMS)[16]. 然而, 这些传统方法在具有高离群点比率的数据中性能有限, 并且在不同场景中缺乏足够的鲁棒性.

近年来, 基于学习的方法取得了显著的突破与进展. 基于“内点的一致性是去除离群点的重要线索”这一前提, 主流基于学习的方法主要关注如何有效地捕获上下文以进行内点一致性挖掘, 如学习寻找优质匹配点方法 (learning to find good correspondences, LFGC)[17]、顺序感知网络 (order-aware network, OANet)[18]、一致性学习网络 (consensus learning network, CLNet)[19]、偏好引导滤波网络 (preference-guided filtering network, PGFNet)[20]. 值得注意的是, 除了从匹配对的直接嵌入中捕获上下文的这些典型方法外, 可学习的运动一致性网络 (learnable motion coherence network, LMCNet)[21] 等一些方法还关注匹配对的潜在运动一致性信息. 运动一致性被证明对于一些传统方法中的离群点去除至关重要[14−16, 22]. 这些传统方法试图从输入中估计场景的运动 (表示为运动场), 这反过来又指导离群点的去除. 然而, 运动场容易受到离群点的污染 (contaminated). 基于卷积神经网络方法 (convolutional neural network-based match, ConvMatch)[23−24] 创新性地嵌入匹配对的运动 (运动向量) 作为输入, 以估计特征空间中的运动场, 使用卷积神经网络 (convolutional neural network, CNN) 来去除运动场中的污染, 同时捕获上下文. 然而, CNN 具有固有的缺点[25]: 1) 卷积操作会使运动场过度平滑, 在复杂场景 (例如大场景差异和深度变化大) 中破坏运动场的不连续性; 2) 卷积核的固定形状和大小限制网络捕捉全局上下文的能力.

在本文中, 针对 CNN 引发的问题, 从不同的角度重新审视了 ConvMatch. ConvMatch 利用图注意力网络 (graph attention network, GAT) 来估计运动场. 具体而言, GAT 在估计过程中自适应地为离群点分配更低的权重, 以更有效地估计局部运动. 实际上, 这种机制有可能过滤掉少数离群点. 然而, 由于输入数据中的离群点比率通常很高, 甚至显著超过内点比率, GAT 无法去除足够多的离群点, 进而难以获得高质量的运动场, 因此需要 CNN 的介入. 基于这一分析, 本文提出一种新的解决方案: 在估计运动场之前, 移除大部分相对容易识别的离群点并捕获上下文, 从而获得一组具有较高内点比率的运动向量. 针对剩余离群点引发的少量噪声, 在估计运动场时利用 GAT 自适应地削弱其影响, 即通过注意力机制为这些离群点分配较低的权重, 减少它们对运动场估计的影响. 这样, 所获得的运动场具有以下优点: 1) 可以更好地保留内点所携带的正确信息, 如果内点分布在具有不同深度或视差的区域, 则估计的运动场自然具有不连续性; 2) 不需要使用 CNN 进行去噪, 从根本上避免 CNN 的缺点.

基于上述想法, 本文提出一种有效的网络, 即运动过滤和调整网络 (motion filtering and adjustment network, MFANet). 具体来说, 设计一个运动过滤 (motion filtering, MF) 模块, 该模块包含多个注意力池化层, 用于捕获上下文并逐步去除离群点. 另外, 设计一个规则化和调整 (regularization and adjustment, RA) 模块, 该模块使用不同尺度下的位置信息来估计和调整运动场.

1) MF 模块的设计

**应对高离群点比率的挑战.** 在特征匹配任务中,

初始匹配离群点 (错误匹配) 的比例往往较高, 有时甚至超过内点 (正确匹配). 尽管 GAT 可通过注意力机制为匹配对分配权重 (理论上可抑制离群点), 但当离群点数量过多时, 其去噪能力会显著下降, 导致估计出的运动场质量受限.

**运动场估计前的预处理.** MF 模块的核心目标是在运动场估计之前对匹配进行预处理, 尽可能滤除易于识别的离群点, 同时保留有效的上下文信息. 该模块可为后续处理 (如 RA 模块) 提供离群点比率显著降低、噪声大幅减少的运动向量, 为后续环节精确估计奠定良好的基础.

**基于注意力机制的逐步过滤与上下文聚合.** MF 模块通过堆叠多个注意力池化层, 逐步聚合局部上下文信息, 并筛选出更可靠的运动向量. 这种层级式结构使模型能够从局部到隐式全局范围内整合信息, 有助于更全面理解匹配对之间的关联.

2) RA 模块的设计动机

**纠正 MF 模块可能带来的过滤偏差.** 尽管 MF 模块能有效去除大量离群点, 但仍可能误滤部分内点. 若直接使用其输出结果进行最终分类, 性能可能并非最优. 因此, RA 模块被设计用于在全局和结构化的信息基础上进行精细化处理.

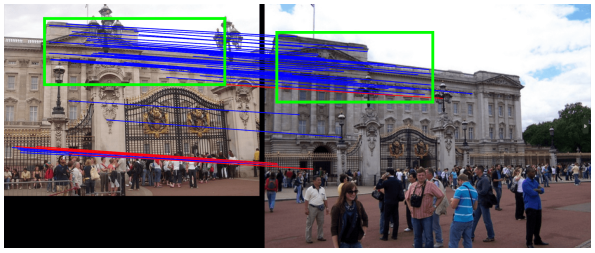
**估计更精确的运动场.** 运动场反映了图像对中对应点运动向量的空间分布. RA 模块旨在生成并优化这一分布, 使其更贴合真实运动情况, 尤其适用于复杂场景 (如大视差、深度突变或非刚性运动), 在这些场景中运动场往往呈不连续或分段平滑的特性.

**实现从无序到有序的转换与局部优化.** RA 模块的设计并非简单重复两次使用 ConvMatch 中的 GAT 结构, 而是基于一项有意义的观察: 首先通过网格嵌入将 MF 模块输出的、仍含少量噪声的无序运动向量转换为规则网格上的有序表示, 形成初步运动场; 随后借助更精细的网格嵌入对该运动场进行局部调整, 以捕捉更细致的运动模式, 从而增强对复杂运动的建模能力.

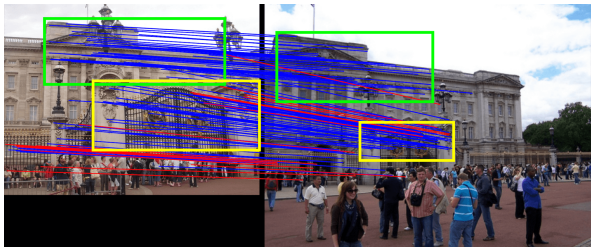
MFANet 通过 MF 模块与 RA 模块的协同配合, 旨在克服现有方法 (如 ConvMatch) 中 GAT 在高离群点比例下性能受限的问题, 以及 CNN 类方法可能导致的运动场过度平滑与全局信息利用不足的局限. MF 模块负责初步筛选与去噪, RA 模块负责精细化运动场估计与优化, 两者共同提升匹配的可靠性与最终分类的准确性. 本文通过 MF 模块和 RA 模块, 构建了运动更新层, 该层是 MFANet 的主要组成部分.

从图 1 可以看出, 该示例中的内点分布于两个不同的区域 (以不同颜色的框标出). U 型网络式匹配 (UNet-like match, U-Match)[26] 与 ConvMatch 仅能识别其中一个区域, 而所提 MFANet 则成功检测出全部两个区域, 并保留更多的内点, 体现其在复杂场景下的鲁棒性. 相较于只能部分保留内点的 ConvMatch, MFANet 在复杂场景中能够保留更完整的内点集合, 从而支持更精确的运动场估计. U-Match 与 ConvMatch 在本图示例场景中表现欠佳, 而本文方法能够克服其局限, 主要原因如下. 图 1 所示场景同时包含重复纹理与显著视角变化, 且存在大量离群点. 在此类极具挑战的场景中:

(a) U-Match



(b) ConvMatch



(c) MFANet (ours)

图 1　通过 U-Match、ConvMatch 和 MFANet 建立的匹配对 (内点和离群点分别用蓝色和红色线标记)

Fig. 1　Matching pairs established by U-Match, ConvMatch, and MFANet (Inliers and outliers are marked with blue and red lines, respectively)

**ConvMatch 的性能受限主要源于其 CNN 的使用.** GAT 模块在高离群点比例下过滤能力有限, 导致部分噪声传递至后续的 CNN 模块. 而 CNN 本身具有的全局平滑特性在处理输入时容易削弱运动场在复杂场景中原本存在的不连续性, 该效应在输入含噪声时会被进一步放大, 最终对运动场估计精度产生不利影响.

**U-Match 的不足主要来自其正交融合模块的**

**设计缺陷.** 该模块通过加权平均池化生成全局特征, 可能导致全局上下文被少数高置信度局部区域主导, 使模型偏向学习特定区域的匹配模式, 而未能充分捕捉图 1 中较小区域的匹配关系.

本文方法通过以下机制提升在此类场景下的性能: 首先, MF 模块在运动场估计前去除大量易识别离群点, 同时保留有效的上下文信息, 为后续处理提供高内点比例的运动向量; 其次, RA 模块在运动场优化阶段利用注意力机制对残余离群点自适应分配低权重, 有效抑制其负面影响, 提升运动场估计的鲁棒性和准确性.

总之, 主要贡献如下:

1) 提出 MF 模块, 以减轻初始匹配对中大多数离群点的负面影响. MF 模块使用多个注意力池化层去除离群点, 并捕获有效的上下文.

2) 提出 RA 模块, 用于生成和调整有序运动向量, 从而表示潜在的运动场. RA 模块对无序运动向量进行规则化, 并引入额外的位置信息来调整有序运动向量, 以增强对复杂运动的表示.

3) 基于 MF 模块和 RA 模块设计一种有效的网络 MFANet, 该网络能够更精确地估计运动场, 进而更有效地去除初始匹配对中的离群点.

## 1 相关工作

在本节中, 简要回顾了过去几十年提出的离群点去除方法, 这些方法一般可分为两类, 即传统方法和基于学习的方法.

### 1.1 传统方法

在传统方法中, RANSAC[9] 是最具代表性的方法, 其迭代地随机选择数据子集, 拟合参数模型, 并根据内点评估其质量. RANSAC 有很多变体, 旨在提高其性能. 例如, Raguram 等[10] 构建了一个综合框架, 该框架兼顾实际应用需求与计算效率. Barath 等[11] 采用一种不依赖于预定义的内点–离群点阈值的方法. 尽管 RANSAC 和这些变体在包含离群点的数据上表现出一定程度的鲁棒性, 但当输入数据的离群点比率超过 50% 时, 这些方法的性能会显著下降. 在实际应用和测试数据集中, 离群点比率通常超过 90%, 因此单独采用 RANSAC 或其变体进行特征匹配并不是最优的.

除 RANSAC 及其变体之外, 多个其他传统方法为去除离群点而设计. 例如, LPM[13] 通过维护局部邻域信息来去除离群点. VFC[14] 通过计算匹配对和向量场之间的兼容性来迭代识别离群点. LAF[15] 使用高斯核卷积对运动场进行采样, 并根据初始匹配对与运动场中典型运动向量之间的一致性来识别离群点. GMS[16] 通过利用基于网格的统计数据来识别可靠的匹配对. 尽管这些方法在去除离群点方面取得了良好进展, 但在应对不同场景时仍然缺乏灵活性[27].

### 1.2 基于学习的方法

近年来, 基于学习的方法在应对各种匹配挑战方面的有效性已得到证实[27]. LFGC[17] 是首个利用深度学习技术进行离群点去除的方法, 该方法基于多层感知器 (multi-layer perceptron, MLP) 设计一个深度网络来识别内点/离群点, 并以端到端的方式进行训练. 然而, 该方法存在一些不足之处: 每个匹配对单独应用上下文归一化 (context normalization, CN), 因此无法捕获邻域内的运动相似性. 此外, CN 通过计算特征图的均值和方差来捕获全局上下文, 忽略了不同匹配对之间的多样性和潜在关系. 这些缺点限制了 LFGC 的性能.

相邻内点之间的一致性是识别内点的关键线索, 而一致性的学习依赖于内点的上下文. 基于这一前提, 许多后续研究设计了复杂的模块来更好地探索内点的上下文信息. 例如, Zhang 等[18] 设计一个可微池化层和一个可微反池化层来捕获局部上下文, 以及一个顺序感知过滤块来捕获全局上下文. Zhao 等[28] 根据匹配对之间的兼容性分数搜索邻居, 以嵌入可靠的局部上下文. Zhao 等[19] 通过学习匹配对的一致性来逐步去除离群点. Wu 等[29] 整合局部和全局一致性来进行内点识别. 此外, 许多研究利用注意力机制[30] 使网络能够更好地应对各种场景. 因此, 网络可以自适应地关注潜在的内点以捕获上下文. 例如, Sun 等[31] 将注意力机制与 CN 相结合, 增强了 CN 的鲁棒性. Ma 等[32] 设计空间注意力机制和通道注意力机制, 以实现匹配对之间的信息交换. Liu 等[20] 设计一个分组残差注意力模块来选择潜在的内点, 并学习偏好分数过滤离群点. Li 等[26] 构建一个直观的编码器–解码器架构, 通过利用注意力上下文聚合来隐式地聚集上下文. Wang 等[33] 引入层次化上下文聚合框架, 该框架集成多粒度聚类和一致性模块来增强匹配对修剪. Wang 等[34] 采用基于动态图的嵌入层来捕获局部拓扑结构, 这些拓扑结构指导注意力层广泛捕获上下文.

上述大多数基于学习的方法主要侧重于从匹配对的直接嵌入中学习上下文, 而匹配对的潜在运动线索则相对较少受到关注. 早期的工作 LMCNet[21] 注意到运动一致性对于匹配对学习的重要性, 并通过提出的拉普拉斯运动拟合 (Laplacian motion fit-

ting, LMF) 设法捕获. 最近的工作 ConvMatch[23] 创新性地以匹配对的位移 (即运动向量) 作为输入, 以显式地引入运动信息. 然后估计一个运动场, 并使用 CNN 来过滤由离群点引起的噪声. 运动场近似地表示整个场景的运动特征, 并可用于有效地识别与内点对应的运动向量. 然而, CNN 破坏真实场景中运动场的不连续性, 并且其捕获上下文的能力也有限. 尽管 ConvMatch 在引入运动信息方面做出创新, 但是也带来由 CNN 引发的问题. 为消除对 CNN 的依赖, 避免 CNN 的负面影响, MFANet 采用不同的方法: 提出一个运动过滤模块, 在估计运动场之前过滤掉大部分离群点; 对于估计的运动场, MFANet 还使用额外的位置嵌入进行调整.

## 2　算法描述

在本节中, 描述所提出的双视图特征匹配学习方法. 首先对问题进行阐述, 然后叙述所提出的 MFANet 的框架和细节, 最后介绍如何使用网络的输出进行内点预测以及一些实现细节.

### 2.1　问题描述

给定一对描述相似场景的图像, 图像特征匹配的目标是建立两幅图像间可靠的稀疏匹配关系. 本文提出的网络采用基于特征点的间接匹配方式来解决图像特征匹配问题. 首先需要构建一个初始的匹配集合, 具体来说, 使用现有的特征检测器 (例如 SIFT) 在两幅图像中检测关键点及其对应的描述子. 然后使用最近邻匹配器基于某种距离度量 (例如欧几里得距离) 将一张图像中的每个关键点匹配到另一张图像中的最近邻点. 通过这种方式, 可以得到一个初始的匹配集合 $\boldsymbol{C}=\{\boldsymbol{c}_i=(\boldsymbol{x}_i,\ \boldsymbol{y}_i)\}_{i=1}^{N}$, 其中 $\boldsymbol{c}_i$ 是第 $i$ 个匹配对; $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 是第 $i$ 个匹配对的两个关键点, 它们的坐标根据相机内参进行了标准化处理; $N$ 是匹配的数量.

通过上述步骤建立的初始匹配集合通常包含大量离群点, 会严重影响下游视觉任务的性能. 因此, 本文按照 LFGC[17] 的做法, 将两视图特征匹配学习任务视为一个内点/离群点分类问题和一个本质矩阵回归问题. 具体而言, 构建一组假定运动向量集合 $\boldsymbol{M}=\{\boldsymbol{m}_i=(\boldsymbol{x}_i,\ \boldsymbol{d}_i)\}_{i=1}^{N}$, 其中 $\boldsymbol{d}_i=\boldsymbol{y}_i-\boldsymbol{x}_i$ 表示第 $i$ 个匹配中两个关键点的坐标偏移. 接着, 训练一个用于匹配对分类的深度网络, 该网络以 $\boldsymbol{M}$ 作为输入, 并输出一组逻辑值 $\hat{\boldsymbol{Z}}=\{\hat{z}_i\}_{i=1}^{N}$. 最后, 以匹配对及其逻辑值为输入, 使用加权八点算法回归一个本质矩阵. 整个过程可以用下式描述:

$$\begin{cases}\hat{\boldsymbol{Z}}=F_{\phi}(\boldsymbol{M})\\ \hat{\boldsymbol{E}}=g(\boldsymbol{C},\ \hat{\boldsymbol{Z}})\end{cases}\tag{1}$$

其中, $F_{\phi}(\cdot)$ 是具有可学习参数 $\phi$ 的深度网络, $\hat{\boldsymbol{E}}$ 表示预测的本质矩阵, $g(\cdot,\ \cdot)$ 表示加权八点算法[17].

### 2.2　网络结构

MFANet 的网络结构如图 2 所示. 首先, 将输入 $\boldsymbol{M}$ 投影到高维空间, 以便网络能够提取深度特征:

$$\boldsymbol{F}=\{\boldsymbol{f}_i=\mathcal{E}(\boldsymbol{m}_i)\}_{i=1}^{N}\in\mathbf{R}^{N\times D}\tag{2}$$

其中 $\mathcal{E}(\cdot)$ 是一个带有位置嵌入的投影层[30], 用于将每个运动向量映射到 $D$ 维空间. 接着, 与 OANet

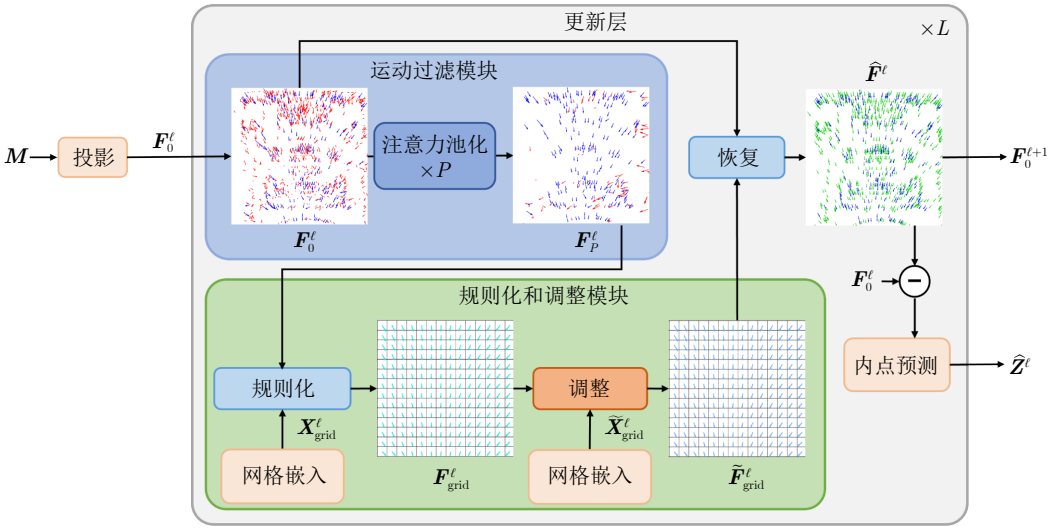



图 2　MFANet 网络结构

Fig. 2　MFANet network structure

和 ConvMatch 类似, 堆叠 $L$ 个运动更新层以迭代更新运动向量. 记第 $\ell$ 个运动更新层的输入为 $\boldsymbol{F}_0^\ell$, 其中 $0 \le \ell \le L-1$ 且 $\boldsymbol{F}_0^0 = \boldsymbol{F}$. 则第 $\ell$ 个运动更新层的迭代过程可表示为:

$$\begin{cases} (\widehat{\boldsymbol{F}}^\ell,\ \widehat{\boldsymbol{Z}}^\ell) = \mathrm{MU}(\boldsymbol{F}_0^\ell) \\ \boldsymbol{F}_0^{\ell+1} = \widehat{\boldsymbol{F}}^\ell \end{cases} \tag{3}$$

其中, MU$(\cdot)$ 表示运动更新层. 每个运动更新层包含两个核心组件: MF 模块和 RA 模块. MF 模块通过堆叠多个注意力池化层构建, 旨在尽可能去除输入数据中的离群点并捕获上下文信息. RA 模块包含规则化和调整过程: 规则化过程生成有序运动向量以表征运动场; 调整过程引入额外位置信息对有序运动向量进行修正, 同时保留复杂运动信息. 最终将有序运动向量恢复为原始分布以进行后续预测. 第 2.3 节和第 2.4 节将详细阐述第 $\ell$ 个运动更新层中的 MF 模块和 RA 模块. 为简化符号, 若无特别说明, 下文将省略所有符号的上标 $\ell$.

### 2.3 运动过滤模块

由于初始匹配集中包含大量离群点, 直接从该集合估计运动场会引入大量噪声, 因此需要对运动场进行噪声过滤. 同时, 这种既要过滤噪声又要保持不连续性的矛盾需求, 显著增加了去噪难度. 为缓解这一问题, MFANet 在规则化处理之前首先过滤 $\boldsymbol{F}_0$ 中的大部分离群点对应的运动向量. 具体而言, 受 U-Match[26] 启发, 本文提出一种基于注意力池化的运动过滤模块. 该模块的主要目标是通过渐进式上下文聚合和离群点过滤, 显著降低 $\boldsymbol{F}_0$ 中离群点对应运动向量的比例, 从而为后续运动场估计提供更纯净的输入数据.

在 MF 模块中, 共堆叠了 $P$ 个注意力池化层. 第 $p$ 个注意力池化层的输入记为 $\boldsymbol{F}_p$, 其中 $\boldsymbol{F}_p$ 包含 $N_p$ 个运动向量 ($0 \le p \le P-1$, 且初始条件为 $N_0 = N$). 该层的输出 $\boldsymbol{F}_{p+1}$ 将作为下一层注意力池化层的输入.

**注意力池化层.** 如图 3 所示, 在第 $p$ 个注意力池化层中, 首先采用 PointCN 模块[17] 和线性层 $\Gamma$ 评估每个运动向量的置信度. 随后根据采样率 $r_p \in [0,\ 1]$ 保留 $k_p$ 个置信度最高的运动向量. 其中, $k_p = N_p \times r_p$ 表示第 $p$ 层需保留的运动向量数量, 同时也作为下一注意力池化层输入的运动向量数量 (即 $N_{p+1} = k_p$). 该采样过程可表述为:

$$\begin{cases} \boldsymbol{s}_p = \mathrm{Sigmoid}(\Gamma(\mathrm{PointCN}(\boldsymbol{F}_p))) \in \mathbf{R}^{N_p} \\ \widetilde{\boldsymbol{F}}_p = \text{top-}k(\boldsymbol{F}_p,\ \boldsymbol{s}_p,\ r_p) \in \mathbf{R}^{k_p \times D} \end{cases} \tag{4}$$

式中, $\boldsymbol{s}_p$ 表示 $N_p$ 个运动向量的置信度; top-$k$ 为筛选函数, 返回 $\boldsymbol{F}_p$ 中置信度前 $k_p$ 个的运动向量构成 $\widetilde{\boldsymbol{F}}_p$.

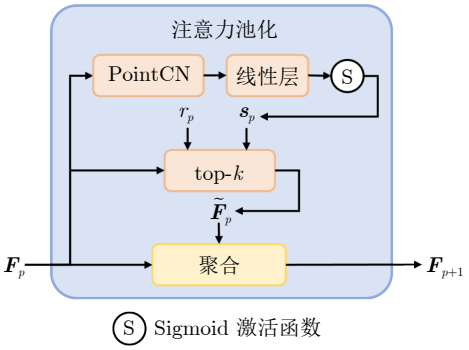



图 3 单个注意力池化层

Fig. 3 Single attention pooling layer

通过式 (4) 操作后, 一部分运动向量被丢弃. 然而, 由于网络可能出现误判, 部分内点对应的运动向量可能被错误丢弃, 导致丢失一些有价值的上下文信息. 因此, 采用注意力机制[30] 将输入的局部上下文信息聚合到保留子集 $\widetilde{\boldsymbol{F}}_p$ 中:

$$\boldsymbol{F}_{p+1} = \mathrm{Cgg}(\widetilde{\boldsymbol{F}}_p,\ \ \boldsymbol{F}_p) \tag{5}$$

其中 Cgg$(\cdot,\ \cdot)$ 表示上下文聚合运算, 其通用形式为:

$$\begin{cases} \boldsymbol{Q} = \mathcal{P}_q(\boldsymbol{X}),\ \boldsymbol{K} = \mathcal{P}_k(\boldsymbol{Y}),\ \boldsymbol{V} = \mathcal{P}_v(\boldsymbol{Y}) \\ \mathrm{Cgg}(\boldsymbol{X},\ \boldsymbol{Y}) = \boldsymbol{X} + \mathrm{MLP}([\boldsymbol{X},\ \mathrm{MHA}(\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V})]) \end{cases} \tag{6}$$

其中, $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别为多头注意力 MHA 运算的查询 (query)、键 (key) 和值 (value); $\mathcal{P}_q(\cdot)$、$\mathcal{P}_k(\cdot)$ 和 $\mathcal{P}_v(\cdot)$ 分别为对应的线性投影层; $[\cdot,\ \cdot]$ 表示连接操作.

经过 $P$ 次注意力池化操作后, 得到包含较少离群点对应运动向量的集合 $\boldsymbol{F}_P$. 值得注意的是, 注意力池化在本方法中具有双重作用, 既实现了上下文聚合, 也完成了离群点去除, 这种双重机制有助于后续更准确地估计运动场.

### 2.4 规则化和调整模块

**运动向量规则化.** 经过运动过滤模块后得到 $\boldsymbol{F}_P$ 中的运动向量分布呈现无序且不均匀的特性, 因此无法有效反映场景的整体运动特征. 传统方法通常需要设计手工规则化器, 通过插值处理将无序运动向量转换为规则运动场[14−15]. 虽然这种运动场为场景运动特征提供统一表征形式, 但手工设计的方式难以适配基于学习的高维数据. 为此, 效仿 ConvMatch[23] 的做法, 采用图注意力网络将无序运动向量转换为有序表征, 以描述高维空间中的运动场. 具体实现时, 将归一化的二维空间划分为 $h \times h$

个非重叠网格单元, 并将无序运动向量采样至这些网格中, 其数学表达为:

$$\begin{cases} \boldsymbol{X}_{\text{grid}} = \text{Up}(\boldsymbol{x}_{\text{grid}}) \in \mathbf{R}^{h\times h\times D} \\ \boldsymbol{F}_{\text{grid}} = \left\{\boldsymbol{f}_{i,\,j}^{\text{grid}}\right\} = \mathcal{G}\left(\boldsymbol{F}_P,\ \boldsymbol{X}_{\text{grid}}\right) \in \mathbf{R}^{h\times h\times D} \end{cases} \tag{7}$$

其中, $\text{Up}(\cdot)$ 表示将输入数据嵌入高维空间的多层感知机; $\boldsymbol{x}_{\text{grid}}$ 是带有位置编码的网格集合; $\boldsymbol{f}_{i,\,j}^{\text{grid}}$ 表示有序运动向量在目标空间的高维嵌入, $0 \le i, j \le h-1$; $\mathcal{G}(\cdot,\,\cdot)$ 为图注意力网络. 由于运动向量已通过 MF 模块完成预处理, 该规则化过程可生成噪声水平显著降低的运动场表征.

**有序运动向量调整.** 如前所述, 有序运动向量具有显著的位置敏感性. 因此, 引入更精细的位置信息来增强运动场集合 $\boldsymbol{F}_{\text{grid}}$. 具体而言, 构建另一个包含更多网格的二维空间划分 $\widetilde{\boldsymbol{x}}_{\text{grid}}$ 作为额外位置信息:

$$\widetilde{\boldsymbol{X}}_{\text{grid}} = \text{Up}(\widetilde{\boldsymbol{x}}_{\text{grid}}) \in \mathbf{R}^{h'\times h'\times D} \tag{8}$$

其中, 令 $h' > h$, 以确保 $\widetilde{\boldsymbol{X}}_{\text{grid}}$ 比 $\boldsymbol{X}_{\text{grid}}$ 包含更多网格.

为有效利用这一额外位置信息, 直观方案是采用 $\widetilde{\boldsymbol{X}}_{\text{grid}}$ 对 $\boldsymbol{F}_{\text{grid}}$ 进行上采样, 以生成更多有序运动向量. 然而, 这种上采样策略会传播原运动场的估计误差——尤其是当 $\boldsymbol{F}_{\text{grid}}$ 中的有序运动向量存在噪声时, 将导致性能显著下降. 因此, 采用密集网格的位置信息对运动场进行自适应调整: 通过图注意力网络实现运动向量与对应网格位置信息的动态融合, 其数学表达为:

$$\widetilde{\boldsymbol{F}}_{\text{grid}} = \mathcal{G}(\ \widetilde{\boldsymbol{X}}_{\text{grid}},\ \boldsymbol{F}_{\text{grid}}) \tag{9}$$

式中, 运动场中的每个有序运动向量会自适应地从 $\widetilde{\boldsymbol{X}}_{\text{grid}}$ 中对应网格聚合位置信息来调整 $\boldsymbol{f}_{i,\,j}^{\text{grid}}$, 这种动态调整机制有效抑制了误差传播.

值得一提的是, 上式的调整过程与式 (7) 的规则化过程具有相似性, 即二者均实现了位置信息与运动信息的有效融合, 其差异仅在于信息融合的顺序不同. 这种顺序差异导致功能不同: 正则化过程通过插值采样生成运动场, 而调整过程则将更精细的位置信息传播给已存在的运动场. 尽管这种设计看似简洁, 却能有效引导有序运动向量基于更丰富的空间信息进行自适应调整, 从而显著提升对复杂运动模式的表征能力.

### 2.5　内点预测和损失函数

在每个运动更新层中, 效仿文献 [23] 的做法, 获得精确运动场后预测内点. 具体而言, 首先, 通过 $\widehat{\boldsymbol{F}}^{\ell} = \{\widehat{\boldsymbol{f}}_i^{\ell}\} = \mathcal{G}(\widetilde{\boldsymbol{F}}_{\text{grid}}^{\ell},\ \boldsymbol{F}_0^{\ell})$ 将 $\widetilde{\boldsymbol{F}}_{\text{grid}}^{\ell}$ 中的有序运动向量恢复至初始分布. 得到的 $\widehat{\boldsymbol{F}}^{\ell}$ 是一组经过噪声校正的运动向量. 其次, 通过 $\hat{\boldsymbol{z}}_i^{\ell} = \mathcal{P}(\widehat{\boldsymbol{f}}_i^{\ell} - \boldsymbol{f}_0^{\ell})$ 计算 $\widehat{\boldsymbol{F}}^{\ell}$ 与 $\boldsymbol{F}_0^{\ell}$ 中运动向量间的残差, 其中 $\mathcal{P}(\cdot)$ 是将数据映射为一维逻辑值的 MLP.

对于包含多个运动更新层的深度网络, 效仿文献 [23] 的监督策略: 对每个中间层的输出施加约束, 但仅保留最终层输出用于推理. 为协同优化分类和回归过程, 采用如下混合损失函数:

$$\mathcal{L} = \sum_{\ell=0}^{L-1}(\mathcal{L}_{cls}(\widehat{\boldsymbol{Z}}^{\ell},\ \boldsymbol{Z}) + \lambda\mathcal{L}_{reg}(\widehat{\boldsymbol{E}}^{\ell},\ \boldsymbol{E})) \tag{10}$$

其中, $\lambda$ 为平衡因子; $\widehat{\boldsymbol{E}}^{\ell}$ 是根据 $\widehat{\boldsymbol{Z}}^{\ell}$ 估计的本质矩阵; $\boldsymbol{E}$ 为真实本质矩阵; 分类损失 $\mathcal{L}_{cls}$ 采用二元交叉熵实现; $\widehat{\boldsymbol{Z}}^{\ell} = \{\hat{\boldsymbol{z}}_i^{\ell}\}$; 弱监督标签 $\boldsymbol{Z} = \{\boldsymbol{z}_i\}$ 基于 Sampson 距离[35] (阈值为 $10^{-4}$) 生成; 本质矩阵回归损失 $\mathcal{L}_{reg}$ 定义为:

$$\mathcal{L}_{reg}(\widehat{\boldsymbol{E}},\ \boldsymbol{E}) = \sum_{i=1}^{N}\frac{(\boldsymbol{y}_i^{\text{T}}\widehat{\boldsymbol{E}}\boldsymbol{x}_i)^2}{\|\boldsymbol{E}\boldsymbol{x}_i\|_{[1]}^2 + \|\boldsymbol{E}\boldsymbol{x}_i\|_{[2]}^2 + \|\boldsymbol{E}^{\text{T}}\boldsymbol{y}_i\|_{[1]}^2 + \|\boldsymbol{E}^{\text{T}}\boldsymbol{y}_i\|_{[2]}^2} \tag{11}$$

式中 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 为匹配对 $\boldsymbol{c}_i$ 的关键点, $||\boldsymbol{v}||_{[i]}$ 表示向量 $\boldsymbol{v}$ 的第 $i$ 个元素.

### 2.6　实现细节

为与 ConvMatch[23] 开展公平对比, MFANet 采用 6 层堆叠的运动更新层 (即 $L = 6$), 其数量与文献 [23] 的主干网络保持一致. 运动更新层中的每个 MF 模块包含两个采样比例均为 0.5 的池化层. 在 RA 模块中, 将规则化的 $h$ 值设为 16, 调整阶段的 $h'$ 值设为 24. 每个多头注意力机制配置 4 个注意力头. 所有实验均在 Ubuntu 20.04 系统下基于 NVIDIA RTX3090 GPU 开展, 其余配置均与文献 [23] 保持一致.

## 3　实验与分析

本节系统评估了 MFANet 在相对位姿估计与离群点去除任务中的表现. 此外, 还通过消融实验分析 MF 模块和 RA 模块的有效性, 从而验证所提方法相比基准算法的优势所在.

### 3.1　数据集和评估指标

针对相对位姿估计与离群点去除任务, 本实验采用不同的评估指标, 并在四个公开的大型室内外场景数据集上对所提出网络进行了训练与测试验证.

YFCC100M 数据集[36] 包含约 1 亿张雅虎从互联网爬取的地标图像 (主要为室外场景), 按场景划分为 72 个图像序列. 数据划分沿用 OANet[18] 的设置: 选取 4 个序列作为测试集, 其余 68 个序列用于模型训练及训练过程中的验证.

SUN3D 数据集[37] 包含大量室内场景的连续视频帧, 所有帧按场景划分为 254 个序列. 其中 239 个序列用于模型训练与验证, 其余 15 个序列作为测试集. 室内场景普遍存在重复结构 (如剧场座椅阵列) 和弱纹理区域 (如光滑桌面) 等复杂特征, 其挑战性显著高于室外场景.

MegaDepth 数据集[38] 基于互联网照片对 196 个全球地标进行运动恢复结构与多视角立体重建, 提供 RGB 图像、密集深度图、相机参数及稀疏三维模型. 该数据集包含极端视角和重复纹理等具有挑战性的真实场景条件, 已成为室外图像匹配和相对位姿估计的标准基准. 评估通常采用 MegaDepth-1500[39] 的划分方式, 该划分从“圣心堂”和“圣彼得广场”等场景中采样 1 500 对图像用于测试.

Sintel 数据集[40] 基于开源动画电影《Sintel》创建, 包含洁净版本 (clean) 和最终版本 (final) 两种渲染版本. 最终版本添加了强烈的大气效果、运动模糊和相机噪声, 对现有方法构成严峻挑战.

**相对位姿估计.** 采用最大平移误差和旋转误差在阈值 (5°、10°、20°) 下的累积误差曲线积分面积 (area under the curve, AUC) 作为核心评估指标. 另外, 还采用角度误差的平均精度 (mean average precision, mAP) 作为补充评估指标. 实验结果包含 RANSAC 与加权八点算法处理后的数据对比, 旨在验证所提方法对不同后处理算法的鲁棒性.

**离群点去除.** 使用分类指标评估性能, 包括准确率 (precision, Pr)、召回率 (recall, R) 和 F 得分 (F-score, F), 以评估所提出网络对初始匹配集分类的准确性. 具体来说, 准确率定义为预测的内点中真实内点的占比; 召回率定义为正确预测内点数占真实内点总数的比例; F 得分综合考虑准确率和召回率两个指标, 用于衡量算法的整体分类性能.

### 3.2 对比实验

将所提出的 MFANet 与传统的方法和基于学习的离群点去除方法进行对比, 其中传统方法包括 RANSAC[9], GMS[16], LPM[13]; 基于学习的方法包括 OANet[18], CLNet[19], 多重稀疏语义动态图网络 (multiple sparse semantics dynamic graph network, MS$^2$ DGNet)[41], PGFNet[20], U-Match[26], ConvMatch[23], 局部一致性变换器方法 (local consensus transformer, LCT)[34] 和共识引导的层次化上下文聚合网络 (consensus-guided hierarchical context aggregation network, CHCANet)[33]. 除非另有说明, 否则在所有实验中, 将具有正逻辑值 ($\hat{z}_i > 0$) 的初始匹配对识别为内点, 因此在评估期间, 将内点判定阈值设置为 0.

**相机姿态估计.** 相机姿态估计任务的对比结果如表 1 ~ 4 所示. 对比的方法包含传统的特征匹配方法和基于深度学习的特征匹配方法两类. 表 1 和表 2 展示了在 YFCC100M 上, 使用 RANSAC 和不使用 RANSAC 两种情况下的比较结果. 可以看到, 在不同的误差阈值下, MFANet 的性能均优于其他方法. 相比于 ConvMatch 和 LCT, MFANet 在使用 RANSAC 进行后处理时, 在 AUC @5° 上分别领先 2.90 和 3.44 个百分点; 在不使用 RANSAC 的情况下分别领先 5.95 和 10.47 个百分点. 即使与性能优于 ConvMatch 的 U-Match 相比, MFANet 仍然具有一定优势. 此外, 与 CHCANet 相比, MFANet 依旧展现出更优的性能.

从表 3 和表 4 可以看出, 在 SUN3D 数据集上, 一些较新的方法 (如 U-Match、ConvMatch 以及所提出的 MFANet) 在不使用 RANSAC 时能够取得较为显著的性能优势. 特别是 MFANet, 是在所有对比方法中唯一一个 AUC @5° 超过 9.00% 的方法. 然而, 在使用 RANSAC 作为额外的后处理步骤时, 这些较新的方法相比于以往的方法, 性能提升更加有限. 这是由于相比于 YFCC100M, SUN3D 数据集中的图像包含更多的弱纹理、重复结构、遮

表 1 在 YFCC100M 数据集上基于 RANSAC 的相机姿态估计比较结果 (%)

Table 1 Comparative results of RANSAC-based camera pose estimation on the YFCC100M dataset (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| RANSAC | 3.63 | 9.00 | 18.32 |
| GMS | 12.12 | 22.78 | 35.27 |
| LPM | 15.18 | 27.30 | 40.93 |
| OANet | 28.07 | 46.22 | 62.99 |
| CLNet | 31.26 | 51.50 | <u>69.11</u> |
| MS$^2$ DGNet | 31.01 | 50.80 | 68.38 |
| PGFNet | 30.59 | 49.28 | 66.07 |
| U-Match | <u>33.54</u> | <u>52.41</u> | 68.96 |
| ConvMatch | 31.53 | 51.07 | 68.11 |
| LCT | 30.99 | 50.65 | 68.09 |
| CHCANet | 30.36 | 49.94 | 67.48 |
| ours | **34.43** | **54.05** | **70.46** |

表 2　在 YFCC100M 数据集上且未采用 RANSAC 的相机姿态估计比较结果 (%)

Table 2　Comparative results of camera pose estimation without using RANSAC on the YFCC100M dataset (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| OANet | 15.94 | 35.90 | 57.05 |
| CLNet | 24.56 | 44.64 | 63.58 |
| MS² DGNet | 18.59 | 40.48 | 62.63 |
| PGFNet | 20.85 | 42.21 | 62.19 |
| U-Match | 30.86 | 52.05 | 69.65 |
| ConvMatch | 25.31 | 47.26 | 66.53 |
| LCT | 20.79 | 42.16 | 63.00 |
| CHCANet | 20.80 | 42.60 | 63.31 |
| ours | **31.26** | **53.37** | **71.06** |

表 3　在 SUN3D 数据集上基于 RANSAC 的相机姿态估计比较结果 (%)

Table 3　Comparative results of RANSAC-based camera pose estimation on the SUN3D dataset (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| RANSAC | 0.96 | 3.29 | 8.66 |
| GMS | 3.60 | 9.02 | 17.68 |
| LPM | 4.82 | 12.30 | 23.62 |
| OANet | 6.81 | 17.14 | 32.46 |
| MS² DGNet | 7.18 | 17.91 | 33.72 |
| PGFNet | 6.80 | 17.22 | 32.53 |
| U-Match | 7.11 | 17.82 | 33.66 |
| ConvMatch | 7.03 | 18.12 | 34.22 |
| LCT | 7.38 | 18.57 | **34.77** |
| CHCANet | 7.22 | 17.78 | 33.38 |
| ours | **7.40** | **18.61** | 34.76 |

表 4　在 SUN3D 数据集上且未采用 RANSAC 的相机姿态估计比较结果 (%)

Table 4　Comparative results of camera pose estimation without using RANSAC on the SUN3D dataset (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| OANet | 5.92 | 16.90 | 34.33 |
| MS² DGNet | 6.31 | 17.76 | 35.78 |
| PGFNet | 5.60 | 16.35 | 33.46 |
| U-Match | 8.05 | 20.81 | 38.71 |
| ConvMatch | 8.39 | 21.75 | 40.01 |
| LCT | 5.83 | 16.52 | 33.42 |
| CHCANet | 6.93 | 18.65 | 36.48 |
| ours | **9.15** | **22.91** | **41.06** |

挡等挑战性因素, 从而算法更容易将离群点错误地预测为内点. 换句话说, 算法预测的内点中, 真实离群点的比率更高. 如此一来, 对高离群点比率敏感的 RANSAC 在有限迭代次数内就难以估计出更准确的本质矩阵.

此外, 为全面比较 MFANet 和其他方法的性能, 额外对比了在两个数据集上的 mAP 指标, 结果如表 5 和表 6 所示. 可以看到, MFANet 在 mAP 指标下依然表现出了最优的性能, 并且性能优势更为明显. 在 YFCC100M 数据集上, 相比于 ConvMatch, 所提出的 MFANet 在使用 RANSAC 进行后处理时, 在 mAP @5° 上领先 3.07 个百分点, 在不使用 RANSAC 的情况下领先 4.98 个百分点. 同样地, 在 SUN3D 数据集上, MFANet 分别领先 0.57 个百分点和 1.02 个百分点.

此外, 更可靠的特征提取算法能够得到更多的特征点和更加鲁棒的描述子, 因此, 特征匹配的最终质量同样取决于特征提取算法的可靠性. 表 7 给出当基于不同的特征提取算法 (SIFT 和 RootSIFT) 构建初始匹配集合时, 所提出的 MFANet 和部分对比方法的性能比较结果. 可以看到, 当使用基于 RootSIFT 构建的初始匹配集合时, 大部分方法都取得了更好的性能. 其中, MFANet 依然在所有方法中具有最佳的性能表现.

**离群点去除.** 离群点去除任务上的性能对比同样是在 YFCC100M 和 SUN3D 数据集上进行. 如表 8 所示, MFANet 和现有的一些优秀方法在准确度、召回率和 F 得分上进行对比.

表 5　在 YFCC100M 数据集上进行相机姿态估计时, 使用/不使用 RANSAC 两种情况下的比较结果 (%)

Table 5　Comparative results of camera pose estimation with and without using RANSAC on the YFCC100M dataset (%)

| 方法 | mAP | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| RANSAC | 9.08/— | 14.28/— | 22.80/— |
| GMS | 26.30/— | 34.59/— | 40.43/— |
| LPM | 28.78/— | 37.55/— | 47.47/— |
| OANet | 52.35/39.23 | 62.10/53.76 | 72.08/67.63 |
| CLNet | 58.60/42.98 | 68.99/53.13 | 78.98/63.47 |
| MS² DGNet | 57.25/45.03 | 68.10/59.90 | 77.97/73.99 |
| PGFNet | 55.23/47.95 | 65.49/61.03 | 75.16/72.98 |
| U-Match | 59.70/60.33 | 69.43/71.09 | 78.41/80.28 |
| ConvMatch | 58.58/57.05 | 68.44/68.79 | 78.04/78.79 |
| LCT | 57.48/47.43 | 67.43/60.91 | 77.29/74.04 |
| CHCANet | 56.53/46.92 | 67.31/61.06 | 77.41/74.16 |
| ours | **61.65/62.03** | **71.26/72.98** | **80.06/81.87** |

表 6 在 SUN3D 数据集上进行相机姿态估计时, 使用/不使用 RANSAC 两种情况下的比较结果 (%)

Table 6 Comparative results of camera pose estimation with and without using RANSAC on the SUN3D dataset (%)

| 方法 | mAP | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| RANSAC | 2.86/— | 5.61/— | 11.22/— |
| GMS | 10.58/— | 16.63/— | 21.59/— |
| LPM | 12.16/— | 19.08/— | 28.93/— |
| OANet | 17.34/16.35 | 26.67/35.70 | 39.54/42.19 |
| CLNet | 17.70/9.96 | 27.61/18.56 | 40.99/31.70 |
| MS$^2$ DGNet | 17.98/17.35 | 27.66/28.71 | 40.98/43.99 |
| PGFNet | 17.44/15.49 | 26.70/26.27 | 39.57/41.41 |
| U-Match | 18.01/21.40 | 27.85/32.68 | 41.13/47.12 |
| ConvMatch | 18.74/22.66 | 28.77/34.53 | 42.12/49.04 |
| LCT | 17.31/16.04 | 27.08/26.62 | 40.05/41.26 |
| CHCANet | 17.33/17.80 | 27.16/28.88 | 40.37/43.41 |
| ours | **19.31/23.68** | **29.12/35.34** | **42.31/49.69** |

表 7 在 YFCC100M 数据集上使用不同特征提取算法的比较结果 (%)

Table 7 Comparative results of various feature extraction methods on the YFCC100M dataset (%)

| 特征 | 方法 | AUC | | |
|---|---|---|---|---|
| | | @5° | @10° | @20° |
| SIFT | OANet | 28.07/15.94 | 46.22/35.90 | 62.99/57.05 |
| | PGFNet | 30.59/20.85 | 49.28/42.21 | 66.07/62.19 |
| | U-Match | 33.54/30.86 | 52.41/52.05 | 68.96/69.65 |
| | ConvMatch | 31.53/25.31 | 51.07/47.26 | 68.11/66.53 |
| | ours | **34.43/31.26** | **54.05/53.37** | **70.46/71.06** |
| RootSIFT | OANet | 29.74/17.69 | 48.77/38.18 | 65.62/59.05 |
| | PGFNet | 31.25/22.78 | 50.72/44.76 | 67.47/64.21 |
| | U-Match | 33.28/27.86 | 52.84/49.41 | 69.46/67.96 |
| | ConvMatch | 33.16/27.49 | 52.49/49.23 | 68.97/68.03 |
| | ours | **35.25/32.31** | **54.54/54.33** | **70.93/71.71** |

值得注意的是, 表 8 中的数据基于网络预测的逻辑值直接计算得出, 其中将逻辑值大于 0 的匹配对视为预测内点. 在此计算方式下, MFANet 与现有方法相比未表现出明显的性能优势. 然而, 如前述相机姿态估计对比实验所示, MFANet 依然能够实现更准确的相机姿态估计. 为解释这一现象, 图 4 展示了 ConvMatch 与 MFANet 预测逻辑值的对比结果. 该图中, 横轴表示 2 000 个匹配对, 纵轴为网络预测的逻辑值, 绿色点表示真实内点, 橙红色点则表示真实离群点. 从图中可看出, MFANet 所预测的逻辑值在内点与离群点之间具有比 ConvMatch 更显著的区分能力. 为进一步验证该结论, 设置了−2、−1、1 和 2 四个不同的内点阈值进行实验. 表 8 的结果表明, 当阈值提高至 1 和 2 时, MFANet 的 F 得分显著上升. 此外, 图 5 中给出了内点与离群点逻辑值的统计直方图, 可清晰看出 MFANet 倾向于为内点分配较高逻辑值 (逻辑值大于 10 的占比为 49.80%, 而 ConvMatch 中此类情况占比为 0.00%), 同时为离群点分配较低逻辑值 (逻辑值低于 −7.5 的占比达 88.20%, ConvMatch 中同样无此类分布). 这些结果一致表明 MFANet 在

表 8 在 YFCC100M 和 SUN3D 数据集上的离群点去除结果 (%)

Table 8 Outlier removal results on the YFCC100M and SUN3D datasets (%)

| 方法 | 数据集 | | | | | |
|---|---|---|---|---|---|---|
| | YFCC100M | | | SUN3D | | |
| | Pr | R | F | Pr | R | F |
| OANet | 57.54 | 86.64 | 66.94 | 46.91 | 83.69 | 60.12 |
| MS$^2$ DGNet | 59.91 | 87.30 | 71.06 | 47.69 | 84.29 | 60.92 |
| PGFNet | 58.11 | 87.38 | 69.80 | 47.35 | 84.32 | 57.05 |
| U-Match | 60.28 | 90.61 | 72.40 | 47.59 | **85.59** | 61.17 |
| ConvMatch | 60.03 | 89.19 | 71.76 | 47.55 | 84.60 | 60.88 |
| LCT | 59.18 | 87.65 | 70.65 | **48.40** | 83.84 | 61.37 |
| CHCANet | 59.88 | 87.07 | 70.96 | 46.63 | 84.66 | 60.14 |
| ours | **60.97** | **90.72** | **72.93** | 48.02 | 85.19 | **61.42** |
| ours (−2) | 43.14 | 97.91 | 59.89 | 27.40 | 96.96 | 42.73 |
| ours (−1) | 49.43 | 95.90 | 65.24 | 36.31 | 93.12 | 52.25 |
| ours (1) | 72.89 | 84.59 | 78.31 | 60.32 | 76.90 | 67.61 |
| ours (2) | 82.02 | 77.66 | 79.78 | 66.01 | 66.82 | 66.41 |

注: 括号中的值表示用于分类网络预测的逻辑值的内点阈值 (未标注时默认为 0). 当匹配对对应的逻辑值大于该阈值时, 判定该匹配对为内点; 反之, 判定为离群点.

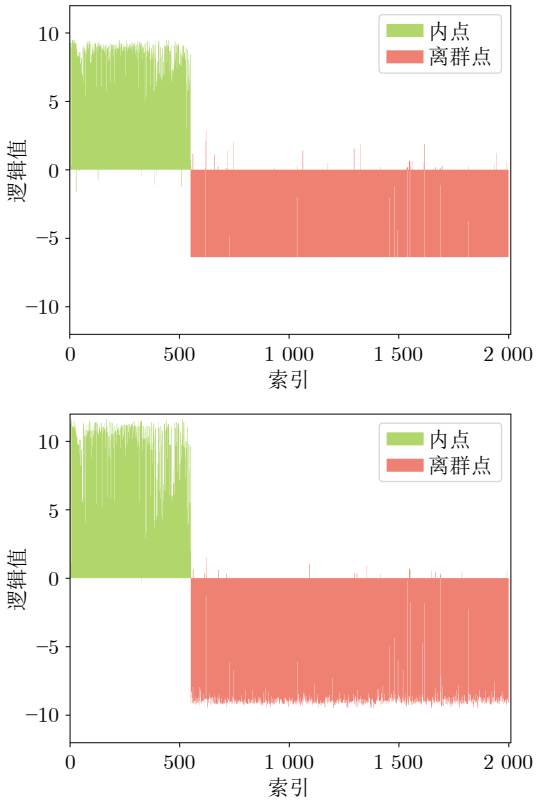



图 4　ConvMatch (上) 和所提出的 MFANet (下) 预测的逻辑值对比

Fig. 4　Comparisons of logical values predicted by ConvMatch (top) and the proposed MFANet (bottom)

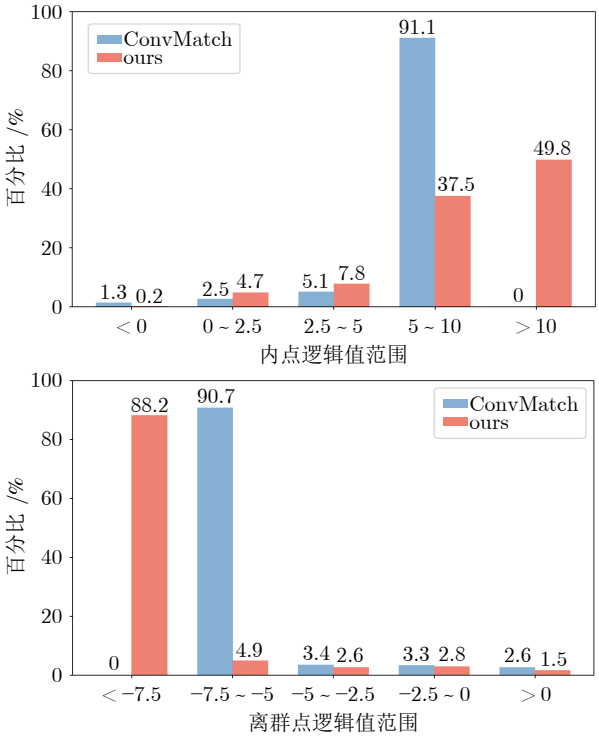



图 5　ConvMatch 和所提出的 MFANet 预测的内点逻辑值 (上) 和离群点逻辑值 (下) 对比

Fig. 5　Comparisons of logical values of inliers (top) and outliers (bottom) predicted by ConvMatch and the proposed MFANet

判别匹配对真实性方面具有更优的置信度区分能力.

根据 YFCC100M 和 SUN3D 数据集上的综合结果, MFANet 展现出先进的性能. 除此之外, 考虑到内点阈值的变化也会影响方法的性能, 图 6 给出 MFANet 和两个效果最好的对比方法 (ConvMatch 和 U-Match) 在不同内点阈值下的性能对比, 实验在 YFCC100M 的测试集上进行. 可以看到, 所提出的 MFANet 在 [−2.5, 3.5] 的内点阈值范围内都取得了不错的性能, 其中取 2 时得到一个较优 F 得分, 这与表 8 的结果一致. 这也得益于 MFANet 能够得到更具区分度的逻辑值分布, 在内点阈值发生变化时具有更低的分类错误率.

可以看到, 虽然总体上 ConvMatch 和 MFANet 预测的逻辑值差别不大, 但所提出的 MFANet 对于真实的内点能够预测出更高的逻辑值, 对于真实的离群点能够预测出更低的逻辑值. 这种更具区分度的权重分布更有利于估计准确的本质矩阵. 为验证这一推论, 本节额外给出了基于本质矩阵计算得到的分类结果. 具体来说, 该计算方式基于网络预测的本质矩阵计算初始匹配集合中所有匹配点对的极线距离, 并将极线距离小于 $10^{-4}$ 的匹配点对作为内点. 这种验证方式在一些利用匹配子集进行相机姿态估计的方法中被使用 (如 CLNet), 这类方法仅输出初始匹配子集的预测逻辑值, 因此必须使用上述方式对完整集合进行验证, 其被称作全尺寸验证 (full-size verification)[19]. 分类结果如表 9 所示, 可以看到, 使用网络预测的本质矩阵对初始匹配进行离群点去除时, 在 YFCC100M 数据集中, 相比于 ConvMatch, 所提出的 MFANet 在准确率、召回率、F 得分上分别领先 2.08、3.03 以及 2.54 个百分点. 而在 SUN3D 数据集上, 虽然准确率提升较小, 但在召回率上有一定提升, 整体 F 得分也略有提高. 总体上说, 使用网络预测的本质矩阵对初始匹配进行离群点去除时, MFANet 取得了最佳的性能.

以上的对比实验展示了 MFANet 和其他方法性能的定量结果. 为直观呈现实际特征匹配效果的差异, 图 7 展示了分别使用 U-Match、ConvMatch 和 MFANet 得到最终匹配的可视化结果. 图中展示了六个来自不同场景的图像对, 其中前三个图像对来自 YFCC100M 的室外场景, 另外三个则来自 SUN3D 的室内场景. 网络预测的内点以直线标出, 其中蓝线为正确的预测, 红线为错误的预测. 可以看到, MFANet 在这些具有挑战性场景中表现出最

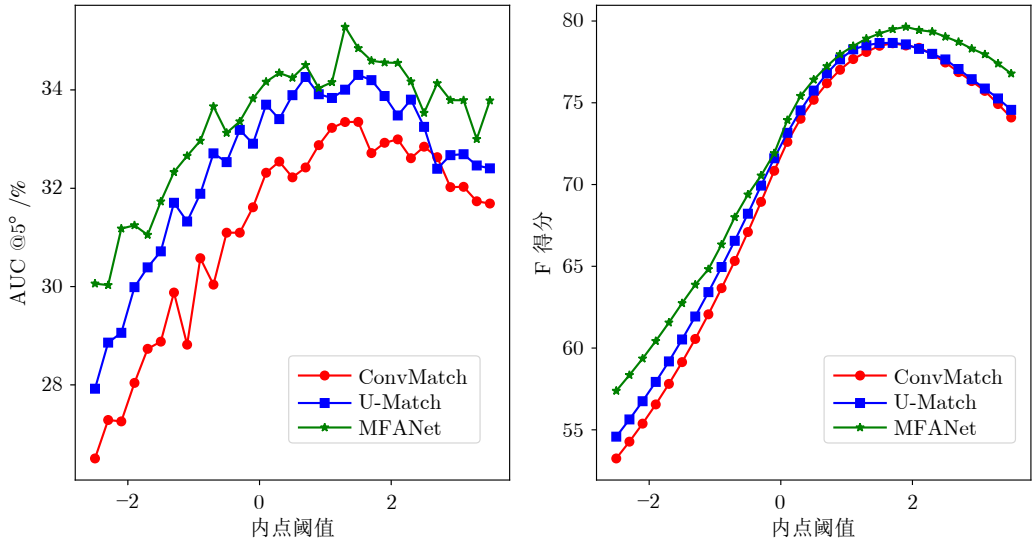



图 6 MFANet、ConvMatch 以及 U-Match 在不同内点阈值下的性能对比

Fig. 6 Performance comparisons of MFANet, ConvMatch, and U-Match at different inlier thresholds

表 9 在 YFCC100M 和 SUN3D 数据集上的离群点去除结果 (%)

Table 9 Outlier removal results on the YFCC100M and SUN3D datasets (%)

| 方法 | 数据集 | | | | | |
|---|---|---|---|---|---|---|
| | YFCC100M | | | SUN3D | | |
| | Pr | R | F | Pr | R | F |
| OANet | 68.04 | 68.41 | 68.22 | 57.66 | 63.11 | 60.26 |
| MS² DGNet | 71.71 | 73.44 | 72.56 | 58.22 | 63.25 | 60.63 |
| PGFNet | 71.00 | 72.26 | 71.62 | 57.84 | 64.00 | 60.76 |
| U-Match | 74.02 | 75.75 | 74.88 | 58.95 | 64.81 | 61.74 |
| ConvMatch | 73.12 | 74.35 | 73.73 | 59.48 | 65.43 | 62.31 |
| LCT | 71.61 | 73.33 | 72.46 | 57.95 | 63.69 | 60.68 |
| CHCANet | 72.05 | 72.93 | 72.49 | 58.43 | 63.86 | 61.02 |
| ours | **75.20** | **77.38** | **76.27** | **59.63** | **65.48** | **62.42** |

注: 本表使用网络预测的本质矩阵计算匹配对的极线距离, 并以极线距离及指定的内点阈值来预测内点.

优的性能. 值得注意的是, 在图中最后一个弱纹理场景中, ConvMatch 和 U-Match 都没能保留正确的内点, 只有 MFANet 成功保留了一部分正确的内点.

### 3.3 效率和泛化实验

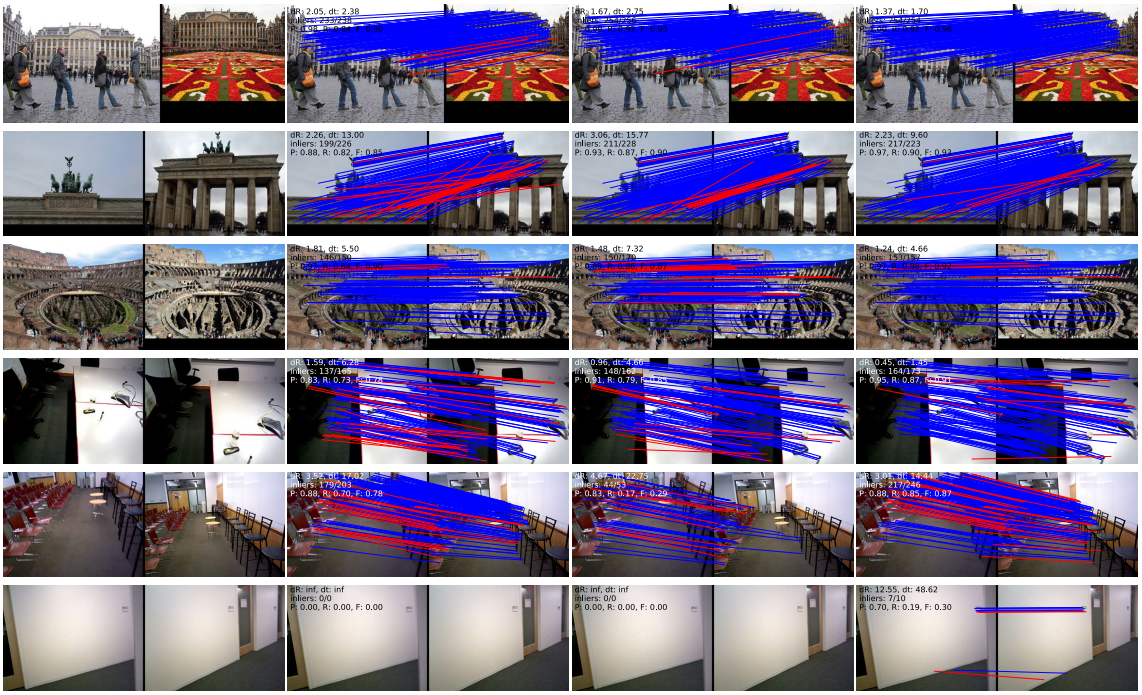
**不同方法的效率.** 表 10 显示了在 YFCC100M 上不同方法的效率与资源消耗比较结果. 与现有方法相比, 本文方法在多项效率指标上展现出均衡而具有竞争力的性能. 从参数量来看, 本方法仅需 5.57 M, 显著低于 U-Match (7.76 M) 和 ConvMatch (7.49 M), 体现出优秀的模型紧凑性和参数利用效率.

在推理时间方面, 本方法仅需 51.1 ms, 与 PGFNet (52.3 ms) 和 U-Match (52.1 ms) 处于同一水平, 展现出良好的实时推理能力. 虽然推理时间略高于 ConvMatch (34.6 ms), 但这一微小的时间代价换来了更强大的特征匹配性能.

在训练效率方面, 本方法需 65 h, 虽高于对比方法, 但这一代价主要来自于 MF 模块的渐进式去噪和 RA 模块的精细运动场优化过程, 这些设计显著提升了模型在复杂场景下的匹配精度和鲁棒性.

MFANet 的 FLOPs 为 9.73 G, 略高于 U-Match (7.48 G) 和 ConvMatch (7.57 G), 反映了本方法在特征提取和运动场建模方面采用了更丰富的计算设计, 这些计算资源投入直接转化为实际匹配任务中更优异的性能表现.

总体而言, MFANet 在参数效率、推理速度与匹配性能之间实现了最佳平衡, 既保持了模型的轻量化特性, 又通过精心设计的计算提升了匹配精度, 是对现有方法的一个有效改进.



(a) 输入图像对
(a) input image pairs

(b) U-Match
(b) U-Match

(c) ConvMatch
(c) ConvMatch

(d) MFANet
(d) MFANet

图 7　U-Match、ConvMatch 和所提出的 MFANet 的可视化结果

Fig. 7　Visualization results of U-Match, ConvMatch, and the proposed MFANet

表 10　在 YFCC100M 上不同方法的效率与资源消耗比较结果

Table 10　Comparative results of the efficiency and resource consumption of different methods on YFCC100M

| 方法 | 参数量 (M) | 推理时间 (ms) | 训练时间 (h) | FLOPs (G) |
|---|---|---|---|---|
| PGFNet | 2.99 | 52.3 | 43 | 3.28 |
| U-Match | 7.76 | 52.1 | 52 | 7.48 |
| ConvMatch | 7.49 | 34.6 | 40 | 7.57 |
| ours | 5.57 | 51.1 | 65 | 9.73 |

**在更大规模数据集 MegaDepth-1500 上的对比.** 实验设置如下: 所有图像的长边统一缩放至 1600 像素; 使用 SIFT 提取初始关键点, 并通过最近邻匹配获得初始对应关系; 随后采用不同算法进行离群点去除, 最终通过 RANSAC 优化得到精确位姿估计. 定量结果如表 11 所示: 在 5°、10°、20° 误差阈值下, 本方法的 AUC 分别达到 41.89%、58.28% 和 72.12%, 相比 CHCANet 提升了 1.10、0.67 和 0.51 个百分点; 与 LCT 和 ConvMatch 方法相比, 本文方法优势更为显著. 尤其在严格阈值 (AUC @5°) 条件下, 本文方法展现出更高的精度提升, 充分验证了其有效性.

**在光流数据集 Sintel 上的对比.** 将所提方法嵌入 PWCNet[42] 框架后 (记为 PWC + ours), 在 Sintel 数据集上的实验结果 (表 12) 表明:

表 11　在 MegaDepth-1500 数据集上且使用 RANSAC 的相机姿态估计比较结果 (%)

Table 11　Comparative results of camera pose estimation with RANSAC on the MegaDepth-1500 dataset (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| ConvMatch | 40.40 | 56.78 | 70.59 |
| LCT | 39.35 | 56.23 | 70.41 |
| CHCANet | 40.79 | 57.61 | 71.61 |
| ours | **41.89** | **58.28** | **72.12** |

表 12　在 Sintel 数据集上光流端点误差的比较结果 (像素)

Table 12　Comparative results of optical flow endpoint error on the Sintel dataset (pixel)

| 方法 | clean | final |
|---|---|---|
| PWCNet | 2.55 | 3.93 |
| PWC + ours | **2.42** | **3.84** |

在 clean 子集上, 光流估计误差从基准方法 PWCNet 的 2.55 像素降至 2.42 像素, 绝对降幅为 0.13 像素; 在 final 子集上, 误差从 PWCNet 的 3.93 像素降至 3.84 像素, 绝对降幅为 0.09 个像素.

以上结果说明, 所提方法能够有效嵌入现有光流网络 PWCNet, 并在其基础上进一步降低光流估计误差, 尤其在不含复杂噪声的 clean 子集上提升

更为显著, 验证了所提方法的有效性和兼容性.

**基于深度学习的局部特征提取算法 LIFT 的性能对比.** 如表 13 所示, 在未使用 RANSAC 后处理的 YFCC100M 数据集上, MFANet 在所有评估维度上的性能均显著优于现有对比方法.

表 13 在 YFCC100M 数据集上使用特征提取算法 LIFT 的比较结果 (不使用 RANSAC) (%)

Table 13 Comparative results of the LIFT feature extraction method on the YFCC100M dataset (without RANSAC) (%)

| 方法 | AUC | | |
|---|---|---|---|
| | @5° | @10° | @20° |
| OANet | 11.42 | 28.85 | 50.26 |
| PGFNet | 14.69 | 33.68 | 54.50 |
| U-Match | 18.85 | 38.80 | 59.38 |
| ConvMatch | 17.75 | 38.57 | 59.86 |
| ours | **22.48** | **43.79** | **64.07** |

在严格阈值 AUC @5° 下, MFANet 达到 22.48%, 较ConvMatch (17.75%) 提升 4.73 个百分点; 在 AUC @10° 和 AUC @20° 阈值下, MFANet 分别达到 43.79%和 64.07%, 相比 ConvMatch (38.57% 和 59.86%) 也呈现稳定优势.

ConvMatch 虽在部分指标上超越 U-Match 等基线, 但其性能仍全面低于 MFANet. 结果表明, MFANet 在不依赖 RANSAC 后处理的条件下, 仍能更有效地实现高精度匹配, 体现出更强的鲁棒性和泛化能力.

### 3.4 消融实验

本节展示所提 MFANet 核心模块的消融研究, 包括运动过滤模块以及规则化和调整模块. 实验选择 ConvMatch 作为基线方法进行对比, 所有实验均在 YFCC100M 上进行训练和测试. 本节首先对运动过滤模块以及规则化和调整模块进行整体的消融实验, 以验证核心模块的有效性, 然后分别对二者的超参数进行具体分析, 探究不同的超参数设置对网络性能的影响.

表 14 展示的实验结果探究了不同模块的组合对网络性能的影响. 其中 MF 表示运动过滤模块, 规则化和调整模块被进一步细分为规则化 REG 和调整 ADJ 两个部分, 而 UPS 表示对运动场进行上采样操作. 从表中可以看出, 运动过滤模块对于性能的提升帮助最大. 从表中的第三行可以看到, 在使用运动过滤模块后, 即使没有引入其他的处理步骤, 网络也能取得良好的性能. 同时可以注意到, 规则化和调整模块在引入运动过滤模块时才能提升网络的性能, 这说明运动场中的过多错误运动信息会干扰调整过程. 此外, 若使用上采样代替调整过程以引入额外的位置信息, 可以得到包含更多有序运动向量的运动场, 但并不能获得更多的性能提升, 这也验证了第 2.4 节的推测 (上采样无法有效提升性能).

表 14 MFANet 在 YFCC100M 上的消融实验结果

Table 14 Ablation experiment results of MFANet on the YFCC100M dataset

| MF | REG | ADJ | UPS | AUC | | |
|---|---|---|---|---|---|---|
| | | | | @5° | @10° | @20° |
| | ✓ | | | 30.43/21.88 | 48.88/42.16 | 65.94/61.65 |
| | ✓ | ✓ | | 29.90/22.42 | 48.72/42.85 | 65.99/62.55 |
| ✓ | | | | 33.18/27.40 | 52.65/49.56 | 69.30/68.22 |
| ✓ | ✓ | | | 34.31/30.23 | 53.59/51.94 | 69.99/69.97 |
| ✓ | ✓ | | ✓ | 34.22/31.03 | 53.31/52.29 | 69.68/69.83 |
| ✓ | ✓ | ✓ | | **34.43/31.26** | **54.05/53.37** | **70.46/71.06** |

注: 本表展示了不同角度误差范围下使用/不使用 RANSAC 作为后处理步骤的比较结果.

所提出的 MF 和 ADJ 模块和 ConvMatch 的网络结构兼容, 因此, 为进一步探究所提出模块的有效性, 本节还将运动过滤模块以及规则化和调整模块中的调整过程集成到了 ConvMatch 的框架中进行实验. 如图 8 所示, 该实验对比了 ConvMatch 及其两个变体的性能. 其中一个变体是在 CNN 之后添加了调整过程 (ConvMatch & ADJ), 另一个变体则是在规则化之前整合了运动过滤模块 (ConvMatch & MF). 实验结果显示, 两个变体相比于 ConvMatch 都具有更好的性能. 如前所述, 调整过

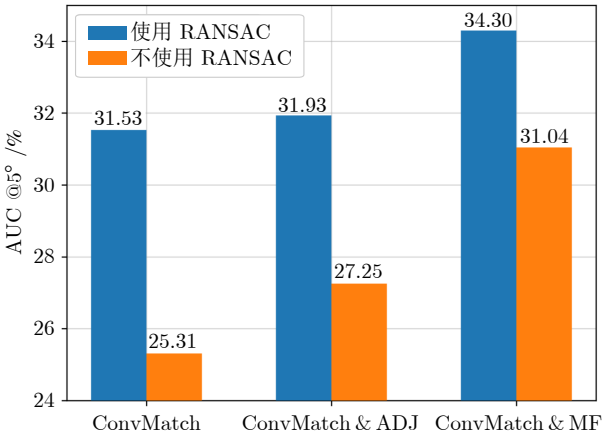



图 8 MF 和 ADJ 集成在 ConvMatch 上的效果

Fig. 8 Performance of ConvMatch with integrated MF and ADJ modules

程依赖输入的运动场的质量, 因此由于 CNN 对运动场进行去噪, 调整过程带来一定程度的性能提升. 然而, 调整过程在 ConvMatch 中的贡献比在 MFANet 中更小, 间接说明相比于使用 CNN, 使用运动过滤模块能够得到更准确的运动场.

**运动过滤模块.** 本节对运动过滤模块的主要超参数进行了分析, 包括注意力池化层的数量 $P$ 和采样率 $r$. 表 15 展示了这两个超参数的分析结果, 该表的实验分为两个部分. 第一部分的实验将采样率固定为 0.50, 探究了注意力池化层数量的影响. 可以看到, 网络的性能随着 $P$ 的增加而提升, 直到 $P=3$ 时性能出现下降. 原因在于过多的池化层会错误地过滤掉一部分的内点, 阻碍了性能的提升. 第二部分的实验目的是研究在保留相同数量的高置信度运动向量的情况下, 池化层数量对网络性能的影响. 因此该部分的实验将池化层数量设置为 1, 采样率设置为 0.25, 如此一来保留的高置信度的运动向量就和第一部分实验的最优配置 ($P=2$, $r=0.5$) 相同. 结果显示, 使用单个池化层时性能出现下降. 这是因为多个池化层能够扩大上下文聚合的范围, 帮助网络更全面地捕获上下文信息, 从而促进离群点的有效过滤. 第三部分的实验目的是研究在池化层数量设置为 2 的情况下, 不同采样率对网络性能的影响. 结果显示, 采样率设为 0.5 仍是最优配置.

表 15　对运动过滤模块的参数分析
Table 15　Parameter analysis of the motion filtering module

| $P$ | $r$ | AUC | | |
|---|---|---|---|---|
| | | @5° | @10° | @20° |
| 0 | — | 29.90 | 48.72 | 65.99 |
| 1 | 0.50 | 32.96 | 52.27 | 69.01 |
| 2 | 0.50 | **34.43** | **54.05** | **70.46** |
| 3 | 0.50 | 33.94 | 53.10 | 69.71 |
| 1 | 0.25 | 32.82 | 52.18 | 68.92 |
| 2 | 0.30 | 32.28 | 51.74 | 68.71 |
| 2 | 0.70 | 33.06 | 52.07 | 68.39 |

注: 本表展示了使用 RANSAC 作为后处理步骤的评估结果.

为展示补充运动过滤模块的效果, 图 9 可视化了使用不同数量的注意力池化层后保留的运动向量, 每个可视化结果都附加显示了离群点占比, 样例取自 YFCC100M 的测试集. 可以看到, 在该样例中, 初始构建的运动向量集合主要由离群点构成. 经过两个注意力池化层的过滤后, 离群点占比显著下降. 这种更加“干净”的运动向量集合将更有利于运动场的准确生成.

**规则化和调整模块.** 对于规则化和调整模块的分析, 主要围绕两个关键超参数 $h$ 和 $h'$ 进行, 这两个超参数分别对应用于规则化过程的网格数量和用于调整过程的网格数量. 图 10 展示了不同的 ($h$, $h'$) 取值对网络性能的影响. 可以看到在不同的参数取值下, 调整过程都在一定程度上提升了性能. 然而在 $h=24$ 时, 若无调整过程, 网络的性能将出现明显下降. 在文献 [23] 中的实验结果表明, $h=24$ 对 ConvMatch 是有益的. 这种差异的出现是因为 MFANet 方法中 MF 模块的输出更“干净”, 但包含的运动向量更少, 不太适合插值到太多的网格中. 幸运的是, 在这种情况下, 调整的引入可以有效地缓解性能下降, 但性能仍然不及 ($h=16$, $h'=24$) 的配置. 因此, MFANet 的最优配置为 ($h=16$, $h'=24$).

当然, 调整过程之所以能够带来性能提升, 也有可能只是单纯地由于额外的 GAT 引入了更多的可学习参数. 为排除这个可能性, 本节对 $h$ 和 $h'$ 进行补充分析. 如表 16 所示, 该表展示了 $h$ 和 $h'$ 更多组合下的性能, 并额外给出网络可学习参数的数

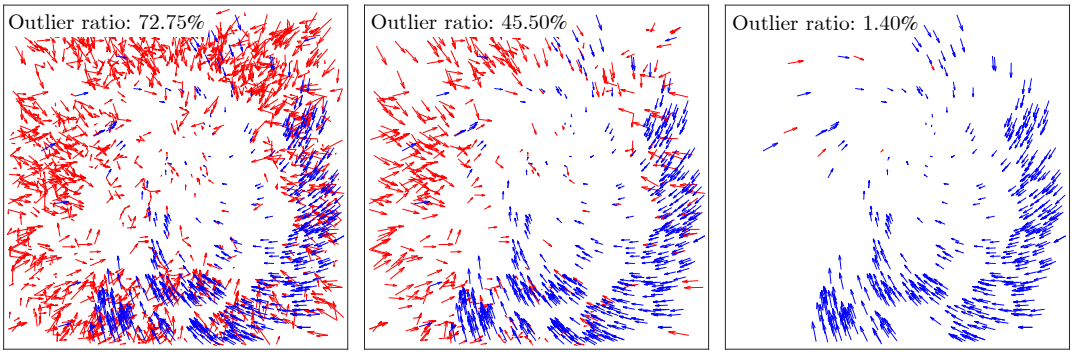



图 9　初始运动向量集合 (左)、使用 1 个注意力池化层 (中)、使用 2 个注意力池化层 (右) 的过滤结果
Fig. 9　Filtering results: Initial motion vector set (left), one attention pooling layer (middle), and two attention pooling layers (right)

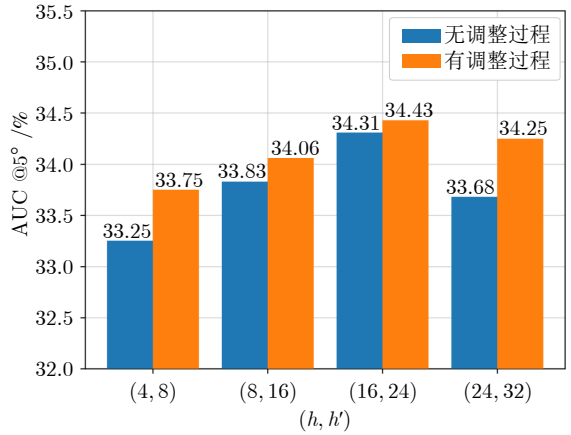



图 10 网格数量对网络性能的影响

Fig. 10 Influence of the number of grids on network performance

表 16 规则化和调整模块不同超参数组合的比较

Table 16 Comparisons of different hyperparameter combinations for the regularization and adjustment modules

| $h$ | $h'$ | 参数量 ($\times 10^6$) | AUC | | |
|---|---|---|---|---|---|
| | | | @5° | @10° | @20° |
| 16 | — | 4.51 | 34.31 | 53.59 | 69.99 |
| 16 | 16 | 5.57 | 33.67 | 53.29 | 70.24 |
| 16 | 24 | 5.57 | **34.43** | **54.05** | **70.46** |
| 24 | 16 | 5.57 | 33.99 | 53.58 | 70.37 |

注: 本表展示了使用 RANSAC 作为后处理步骤的评估结果.

量. 由表中可知, 在较为严格的指标 AUC @5° 和 AUC @10° 下, 除了 ($h = 16$, $h' = 24$) 的配置外, 其他组合都降低了网络性能, 即使它们对网络增加的可学习参数数量相仿. 在 AUC @20° 指标下, 所有组合的性能都有提升, 但 ($h = 16$, $h' = 24$) 的组合依旧是最优的. 这说明更多的可学习参数并不是网络性能提升的原因, 而是规则化和调整模块的设计提高了网络的性能.

### 3.5 失败的样例

在图 7 最后一行的弱纹理场景中, ConvMatch 与 U-Match 均未能保留正确的内点, 而本文所提出的 MFANet 方法成功保留了部分正确内点. 然而, 尽管 MFANet 在一定程度上表现出优于对比方法的性能, 但是在噪声影响下, 若所保留的内点数量不足, 仍可能无法准确估计几何模型. 造成这一局限的原因在于, 与 ConvMatch 和 U-Match 类似, 当前方法主要依赖运动一致性等单一先验, 难以充分学习匹配集合中丰富的多源信息. 在未来的工作中, 可考虑融合多方面的先验知识, 如坐标空间中的几何约束、特征空间中的一致性以及原始图像中的局部语义相似性等, 以增强网络对上下文的全面感知与可靠建模能力.

## 4 结束语

本文提出一个基于运动过滤和调整的图像特征匹配网络 MFANet. 与现有方法不同, MFANet 采用“先去除大部分噪声, 再估计运动场”的思路. 具体来说, 本文设计了一个运动过滤模块, 通过堆叠的注意力池化层过滤初始运动向量集合中的大部分离群点, 得到更加干净的运动向量集合. 在复杂场景中, 这种方式能更好地保留原始运动向量所体现的不同运动模式. 此外, 本文还设计了一个规则化和调整模块, 该模块从干净的运动向量集合中采样一个高质量的运动场, 并引入额外的位置信息来对运动场进行调整, 从而提高运动场估计的准确性. 在室内和室外场景数据集的实验结果均表明了所提出的方法在相机姿态估计任务和离群点去除任务中的优势. 对于网络结构的消融研究也说明了所提出模块的有效性. 总的来说, 相比于现有的多个先进方法, MFANet 取得了显著的性能提升.

## 参考文献


1 Zhang Jun-Ning, Su Qun-Xing, Liu Peng-Yuan, Zhu Qing, Zhang Kai. An improved VSLAM algorithm based on adaptive feature map. *Acta Automatica Sinica*, 2019, **45**(3): 553−565
(张峻宁, 苏群星, 刘鹏远, 朱庆, 张凯. 一种自适应特征地图匹配的改进 VSLAM 算法. 自动化学报, 2019, **45**(3): 553−565)

2 Li Hai-Feng, Liu Jing-Tai. An optimal vanishing point detection method with error analysis. *Acta Automatica Sinica*, 2012, **38**(2): 213−219
(李海丰, 刘景泰. 一种优化的消失点估计方法及误差分析. 自动化学报, 2012, **38**(2): 213−219)

3 Xing X J, Lu Z D, Wang Y Q, Xiao J. Efficient single correspondence voting for point cloud registration. *IEEE Transactions on Image Processing*, 2024, **33**: 2116−2130

4 Peng Z Y, Ma Y, Zhang Y J, Li H, Fan F, Mei X G. Seamless UAV hyperspectral image stitching using optimal seamline detection via graph cuts. *IEEE Transactions on Geoscience and Remote Sensing*, 2023, **61**: Article No. 5512213

5 Lai T T, Sadri A, Lin S Y, Li Z Y, Chen R Q, Wang H Z. Efficient sampling using feature matching and variable minimal structure size. *Pattern Recognition*, 2023, **137**: Article No. 109311

6 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

7 Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 2911−2918

8 Lin S Y, Chen X, Xiao G B, Wang H Z, Huang F R, Weng J. Multi-stage network with geometric semantic attention for two-view correspondence learning. *IEEE Transactions on Image Processing*, 2024, **33**: 3031−3046

9 Fischler M A, Bolles R C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*,

1981, **24**(6): 381−395

10 Raguram R, Chum O, Pollefeys M, Matas J, Frahm J M. USAC: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(8): 2022−2038

11 Barath D, Matas J, Noskova J. MAGSAC: Marginalizing sample consensus. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 10197−10205

12 Lin S Y, Huang F R, Lai T T, Lai J H, Wang H Z, Weng J. Robust heterogeneous model fitting for multi-source image correspondences. *International Journal of Computer Vision*, 2024, **132**(8): 2907−2928

13 Ma J Y, Zhao J, Jiang J J, Zhou H B, Guo X J. Locality preserving matching. *International Journal of Computer Vision*, 2019, **127**(5): 512−531

14 Ma J Y, Zhao J, Tian J W, Yuille A L, Tu Z W. Robust point matching via vector field consensus. *IEEE Transactions on Image Processing*, 2014, **23**(4): 1706−1721

15 Jiang X Y, Ma J Y, Fan A X, Xu H P, Lin G, Lu T, et al. Robust feature matching for remote sensing image registration via linear adaptive filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, **59**(2): 1577−1591

16 Bian J W, Lin W Y, Matsushita Y, Yeung S K, Nguyen T D, Cheng M M. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2828−2837

17 Yi K M, Trulls E, Ono Y, Lepetit V, Salzmann M, Fua P. Learning to find good correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2666−2674

18 Zhang J H, Sun D W, Luo Z X, Yao A B, Zhou L, Shen T W, et al. Learning two-view correspondences and geometry using order-aware network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 5844−5853

19 Zhao C, Ge Y X, Zhu F, Zhao R, Li H S, Salzmann M. Progressive correspondence pruning by consensus learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 6444−6453

20 Liu X, Xiao G B, Chen R Q, Ma J Y. PGFNet: Preference-guided filtering network for two-view correspondence learning. *IEEE Transactions on Image Processing*, 2023, **32**: 1367−1378

21 Liu Y, Liu L J, Lin C, Dong Z, Wang W P. Learnable motion coherence for correspondence pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 3236−3245

22 Ma J Y, Fan A X, Jiang X Y, Xiao G B. Feature matching via motion consistency driven probabilistic graphical model. *International Journal of Computer Vision*, 2022, **130**(9): 2249−2264

23 Zhang S H, Ma J Y. ConvMatch: Rethinking network design for two-view correspondence learning. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington DC, USA: AAAI, 2023. 3472−3479

24 Zhang S H, Ma J Y. ConvMatch: Rethinking network design for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(5): 2920−2935

25 Kang Z, Lai T T, Li Z Y, Wei L F, Chen R Q. PRNet: Parallel reinforcement network for two-view correspondence learning. *Knowledge-Based Systems*, 2025, **310**: Article No. 112978

26 Li Z Z, Zhang S H, Ma J Y. U-Match: Two-view correspondence learning with hierarchy-aware local context aggregation. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: 2023. Article No. 130

27 Ma J Y, Jiang X Y, Fan A X, Jiang J J, Yan J C. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 2021, **129**(1): 23−79

28 Zhao C, Cao Z G, Li C, Li X, Yang J Q. NM-Net: Mining reliable neighbors for robust feature correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 215−224

29 Wu T H, Chen K W. LGCNet: Feature enhancement and consistency learning based on local and global coherence network for correspondence selection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, 2023. 6182−6188

30 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

31 Sun W W, Jiang W, Trulls E, Tagliasacchi A, Yi K M. ACNe: Attentive context normalization for robust permutation-equivariant learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 11283−11292

32 Ma J Y, Wang Y, Fan A X, Xiao G B, Chen R Q. Correspondence attention transformer: A context-sensitive network for two-view correspondence learning. *IEEE Transactions on Multimedia*, 2023, **25**: 3509−3524

33 Wang G, Chen Y F, Wu B. CHCANet: Two-view correspondence pruning with consensus-guided hierarchical context aggregation. *Pattern Recognition*, 2025, **161**: Article No. 111282

34 Wang G, Chen Y F. Two-view correspondence learning with local consensus transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(7): 11861−11874

35 Hartley R, Zisserman A. *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press, 2003.

36 Thomee B, Shamma D A, Friedland G, Elizalde B, Ni K, Poland D, et al. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 2016, **59**(2): 64−73

37 Xiao J X, Owens A, Torralba A. SUN3D: A database of big spaces reconstructed using SFM and object labels. In: Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 1625−1632

38 Li Z Q, Snavely N. MegaDepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2041−2050

39 Sun J M, Shen Z H, Wang Y A, Bao H J, Zhou X W. LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 8918−8927

40 Butler D J, Wulff J, Stanley G B, Black M J. A naturalistic open source movie for optical flow evaluation. In: Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012. 611−625

41 Dai L Y, Liu Y Z, Ma J Y, Wei L F, Lai T T, Yang C C, et al. MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8963−8972

42 Sun D Q, Yang X D, Liu M Y, Kautz J. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8934−8943

**赖桃桃** 闽江学院计算机与大数据学院副教授. 2016 年获得厦门大学计算机科学与技术专业博士学位. 主要研究方向为计算机视觉, 特征匹配, 模型拟合.
E-mail: laitaotao@gmail.com
(**LAI Tao-Tao** Associate professor at the School of Computer and Data Science, Minjiang University. He received his Ph.D. degree in computer science and technology from Xiamen University in 2016. His research interests include computer vision, feature matching and model fitting.)

**张一凡** 福州大学计算机与大数据学院硕士研究生. 主要研究方向为计算机视觉和图像匹配.
E-mail: yifan_fzu@163.com
(**ZHANG Yi-Fan** Master student at the College of Computer and Data Science, Fuzhou University. His research interests include computer vision and image matching.)

**李佐勇** 闽江学院计算机与大数据学院教授. 2010 年获得南京理工大学计算机科学与技术专业博士学位. 主要研究方向为图像处理, 模式识别, 深度学习. 本文通信作者.
E-mail: fzulzytdq@126.com
(**LI Zuo-Yong** Professor at the School of Computer and Data Science, Minjiang University. He received his Ph.D. degree in computer science and technology from Nanjing University of Science and Technology in 2010. His research interests include image processing, pattern recognition, and deep learning. Corresponding author of this paper.)

**肖国宝** 同济大学计算机科学与技术学院教授. 2016 年获得厦门大学计算机科学与技术专业博士学位. 主要研究方向为机器学习, 计算机视觉和模式识别.
E-mail: gbx@tongji.edu.cn
(**XIAO Guo-Bao** Professor at the School of Computer Science and Technology, Tongji University. He received his Ph.D. degree in computer science and technology from Xiamen University in 2016. His research interests include machine learning, computer vision and pattern recognition.)

**林维斯** 新加坡南洋理工大学计算机科学与工程学院教授. 1992 年获得英国伦敦大学国王学院计算机视觉专业博士学位. 主要研究方向为智能图像处理、感知信号建模、视频压缩和多媒体通信.
E-mail: wslin@ntu.edu.sg
(**LIN Wei-Si** Professor at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received his Ph.D. degree in computer vision from King′s College, London University, UK in 1992. His research interests include intelligent image processing, perceptual signal modeling, video compression, and multimedia communication.)

**王菡子** 厦门大学信息学院闽江学者特聘教授. 2004 年获得澳大利亚莫纳什大学计算机视觉专业博士学位. 主要研究方向为计算机视觉.
E-mail: hanzi.wang@xmu.edu.cn
(**WANG Han-Zi** Distinguished professor of Minjiang scholars at the School of Informatics, Xiamen University. He received his Ph.D. degree in computer vision from Monash University, Australia in 2004. His main research interest is computer vision.)