

融合局部和长距离信息的图像特征点提取

唐付林¹⁾, 马朋举²⁾, 段鸿杰²⁾, 郑彦²⁾, 计洪图²⁾, 郭云龙²⁾, 吴毅红^{1)*}

¹⁾ (中国科学院自动化研究所多模态人工智能系统全国重点实验室 北京 100190)

²⁾ (中国石油化工股份有限公司胜利油田分公司 东营 257000)
(yihong.wu@ia.ac.cn)

摘要: 现有的基于学习的图像特征点提取方法忽略了图像中的长距离上下文信息, 在诸如石油生产场景等重复纹理和弱纹理场景中表现不佳, 为油田生产的安全视频监控分析和数字孪生平台的构建带来了较大困难. 为了解决上述问题, 提出一种融合局部和长距离信息的图像特征点提取方法. 首先利用高效的大卷积核捕捉长距离信息, 并将长距离信息与普通卷积捕捉到的局部信息进行有效的融合, 得到判别性更强的图像局部描述子; 为了进一步提升图像特征点的匹配精度, 提出一种全图像范围内自适应最困难描述子三元组构建方法, 将描述子三元组的构造范围由共视区域扩展至整幅图像, 提升训练过程和实际应用场景的相似度. 在 HPatches 数据集上, 所提方法的平均匹配准确率高出对比方法 2.89%. 在 Aachen Day-Night v1.1 数据集上, 所提方法的视觉定位准确率高出对比方法 2.10%. 在 InLoc 数据集上, 所提方法的视觉定位准确率高出对比方法 3.70%. 在富含重复纹理和弱纹理区域的石油生产场景数据集上, 所提方法的平均特征点匹配内点数高出对比方法 20.37%.

关键词: 图像匹配; 局部描述子; 特征提取; 最困难描述子三元组
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00743

Fusing Local and Long-Range Information for Image Feature Point Extraction

Tang Fulin¹⁾, Ma Pengju²⁾, Duan Hongjie²⁾, Zheng Yan²⁾, Ji Hongtu²⁾, Guo Yunlong²⁾, and Wu Yihong^{1)*}

¹⁾ (State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

²⁾ (China Petroleum and Chemical Corporation, SHENGLI OILFIELD, Dongying 257000)

Abstract: Existing learning-based feature point extraction methods neglect long-range context information. Thus, existing learning-based feature point extraction methods perform poorly in repetitive-texture and low-texture scenes such as oil production scenes. This problem poses difficulties for the analysis of safety video surveillance and the construction of digital twin platforms in oil field production. To solve this problem, we propose a novel and efficient feature point extraction network that fuses local and long-range information. Firstly, the proposed network utilizes efficient large-kernel convolutional layers to capture long-range information. Then, the proposed network can extract local descriptors with higher discriminative ability by fusing the long-range information and the local information captured by vanilla convolutional layers. To further enhance the matching accuracy of feature points, we propose a novel method of adaptively constructing hardest descriptor triplets within entire images. The proposed method extends the scope of con-

收稿日期: 2023-11-19; 修回日期: 2024-04-24. 基金项目: 国家自然科学基金(62202468); 中国石油化工股份有限公司“油田生产应急处置应用研究”项目. 唐付林(1992—), 男, 博士, 副研究员, 主要研究方向为基于视觉的同时定位与建图、多传感器融合的实时定位与建图、增强现实; 马朋举(1979—), 男, 硕士, 高级工程师, 主要研究方向为图像匹配; 段鸿杰(1972—), 男, 博士, 教授级高级工程师, 主要研究方向为三维重建; 郑彦(1970—), 男, 学士, 高级政工师, 主要研究方向为相机标定; 计洪图(1983—), 男, 硕士, 高级工程师, 主要研究方向为视觉定位; 郭云龙(1989—), 男, 硕士, 工程师, 主要研究方向为三维重建; 吴毅红(1973—), 女, 博士, 研究员, 博士生导师, 论文通信作者, 主要研究方向为摄像机定标、摄像机定位、图像匹配、三维重建、多视几何学、SLAM、虚拟现实/增强现实.

structing triplets from covisible areas to entire images, which raises the similarity between the training process and real application scenarios. On the HPatches dataset, the mean matching accuracy of the proposed method is 2.89% higher than those of the comparison methods. On the Aachen Day-Night v1.1 dataset, the visual localization accuracy of the proposed method is 2.10% higher than those of the comparison methods. On the InLoc dataset, the visual localization accuracy of the proposed method is 3.70% higher than those of the comparison methods. On the dataset of oil production scenes with repetitive and weak texture regions, the average number of feature point matching inliers of the proposed method is 20.37% higher than those of the comparison methods.

Key words: image matching; local descriptors; feature extraction; hardest descriptor triplet construction

图像特征点提取是图像匹配^[1]、三维重建^[2]、视觉定位^[3]、同时定位与地图构建^[4]等计算机视觉任务的基础步骤。在过去的几十年中,研究者依据先验知识提出许多手工设计的图像特征点提取方法。但在视角和光照变化情况下,手工特征点提取方法的鲁棒性不足。近年来,随着深度学习的发展,研究者也提出了许多基于学习的图像特征点提取方法,其匹配精度高于手工特征点提取方法。

图像特征点由关键点和描述子组成。目前,基于学习的图像特征点提取方法利用卷积神经网络提取图像的局部信息,并依据这些局部信息检测关键点和计算描述子。在重复纹理和弱纹理场景中,现有的基于学习的图像特征点提取方法不能产生具有足够判别性的描述子,从而无法为视觉任务提供足够数量的正确匹配。尤其是对于油田生产的场景,在野外,具有较大的空旷区域,存在典型的弱纹理;在室内,又有大量的管道,存在典型的重复纹理。这些问题为油田生产的安全视频监控分析和数字孪生平台的构建带来了较大的困难。上述问题的成因是,基于学习的图像特征点提取方法采用小卷积核构建网络,其感受野非常有限;在重复纹理和弱纹理情况下,在错误匹配点周围,网络感知的局部区域很可能非常相似,导致错误匹配点具有相似的描述子,造成误匹配。若能增大网络的感受野,则错误匹配点对应的网络感知区域将很有可能不再相似,从而降低错误匹配点的描述子的相似性,有效地避免了误匹配。基于此,本文构建一个轻量的、大卷积核增强的特征提取模块,显著地增大了网络的感受野。该模块中,首先利用大卷积核提取长距离的图像上下文信息;然后利用普通卷积提取图像的局部信息;最后将长距离信息与局部信息进行有效融合,得到判别性更强的描述子。

在描述子学习中,基于学习的图像特征点提

取方法在一对训练图像的共视区域中构建描述子三元组,并基于这些描述子三元组训练描述子模型。然而在实际应用场景中,图像对的共视区域通常是未知的,这就要求描述子不仅能够在图像对的共视区域中进行正确匹配,还要具有在整幅图像范围内进行正确匹配的能力。为了提升描述子训练过程与实际应用场景的相似性,本文提出一种描述子三元组构建方法,将三元组的构建范围由共视区域扩展至整幅图像;还提出一种自适应确定最佳锚点的方法,构建更具有训练价值的描述子三元组,进一步提升图像特征点的匹配精度。

1 相关工作

1.1 手工设计的图像特征点提取方法

手工设计的图像特征点提取方法将一些显著的局部结构确定为关键点,如角点和斑点。SIFT^[5]是一种非常经典的手工特征点提取方法,其基于高斯差分金字塔从多个尺度上提取斑点;Harris^[6]则是一种非常著名的角点检测方法,其依据梯度矩阵检测角点;FAST^[7]通过比较中心像素与周围像素的灰度值来判断中心像素是否为角点;为了更好地保留边缘的位置精度,KAZE^[8]利用非线性扩散滤波方法,从非线性扩散尺度空间中提取关键点。手工设计的图像特征点提取方法利用图像的灰度和梯度信息对关键点周围的图像块进行描述。SIFT^[5]使用梯度方向直方图构建描述子;ORB^[9]先利用灰度质心法计算关键点的主方向,然后计算依据主方向旋转的 BRIEF 描述子^[10];BRIEF 则通过随机比较图像块内的像素灰度值来生成二进制描述子。手工设计的图像特征点提取方法在剧烈光照变化和视角变化情况中的鲁棒性不足,在困难场景,尤其是石油生产的场景中,不能为视觉任

务估计出足够数量的正确匹配。

1.2 基于学习的图像特征点提取方法

基于学习的图像特征点提取方法利用神经网络从图像中提取特征图, 并依据特征图确定关键点和计算描述子。DeTone 等^[11]提出一种自监督的图像特征点学习方法 SuperPoint, 首先用模拟生成的规则图像和人工标注信息训练一个初始关键点检测器, 然后利用单应变换和初始关键点检测器对真实图像进行关键点标注, 最后利用单应变换生成模拟图像对进行图像特征点学习; D2-Net^[12]则依据空间局部显著性和通道显著性从特征图上检测关键点, 利用三元损失函数训练网络的描述子提取能力; ASLFeat^[13]改进了 D2-Net 的关键点提取策略, 通过融合不同尺度上的特征图信息进行更加精确的关键点检测; R2D2^[14]则依据可重复性和匹配可靠性进行鲁棒的关键点检测; ALIKE^[15]使用神经重投影误差学习描述子; Dusmanu 等^[16]提出一种将不同种类的描述子转化为可匹配的中间描述子的跨描述子定位与建图方法; Shen 等^[17]提出一个感知尺度的特征点提取网络和一个圆环二阶相似度正则化损失函数; Fan 等^[18]提出一种基于对抗学习的方法, 增强 ASLFeat^[13]在昼夜光照变化情况中的鲁棒性; Wang 等^[19]提出一种不使用负样本的自监督特征点学习方法; Wang 等^[20]则提出一个用于增强已有描述子判别性的神经网络。虽然基于学习的图像特征点提取方法的匹配精度高于手工设计的图像特征点提取方法, 但上述基于学习的方法均采用卷积神经网络提取图像特征, 由于卷积神经网络只能感知局部信息, 因此在重复纹理和弱纹理情况下, 上述基于学习的方法无法提取出具有足够判别性的描述子, 不能为后续视觉应用建立足够数量的正确匹配。为了解决该问题, 本文提出一个轻量的大卷积核增强的特征提取模块, 通过融合局部和长距离信息, 提取出具有更强判别性的描述子。鉴于 ASLFeat^[13]仅在一对训练图像的共视区域内构建描述子三元组, 为了构建更有价值的描述子三元组, 本文将描述子三元组的构建范围扩展至整幅图像, 以提升描述子训练过程与实际应用场景的相似性; 本文还提出一种自适应选择最佳锚点的方法, 进一步提升了描述子的判别性。

2 本文方法

2.1 高效的大卷积核增强的特征提取模块

为了提升基于学习的图像特征点提取方法的

匹配性能, 本文提出一个大卷积核增强的特征 (large-kernel-convolution-reinforced feature, LKCRF) 提取模块。该模块能够融合局部和长距离上下文信息, 从而提取出具有更强判别性的描述子。本文构建的图像特征点提取网络 LKCRF-Net 的结构如图 1 所示。

给定一幅输入图像 I , H 和 W 分别表示该图像的高度和宽度, 本文使用多个连续的卷积层、池化层和 ReLU 激活函数提取图像的多尺度特征 F_1 , F_2 和 F_3 。为了进行精确的多尺度关键点检测, 使用 3×3 卷积、Sigmoid 函数和双线性插值操作构建了 3 个关键点检测头, 它们依据 F_1 , F_2 和 F_3 估计出 3 个尺度的关键点热图 S_1 , S_2 和 S_3 。为了融合多个尺度的关键点信息, 将 S_1 , S_2 和 S_3 进行加权求和, 得到最终的关键点热图 $S = \omega_1 S_1 + \omega_2 S_2 + \omega_3 S_3$; 其中, ω_1 , ω_2 和 ω_3 表示权重, $\omega_1 = \omega_2 = \omega_3 = 1/3$ 。LKCRF-Net 输出的多尺度关键点热图如图 2 所示。可以看出, 第 1 尺度关键点热图 S_1 更关注图像的细节结构信息, 但对边缘等不能被稳定匹配的区域^[5]过于敏感; 第 2 尺度关键点热图 S_2 和第 3 尺度关键点热图 S_3 可有效地抑制网络对于边缘等不鲁棒细节结构的响应值, 并进一步强调角点和斑点等能够被稳定匹配的区域。

在描述子提取方面, 本文致力于融合局部和长距离上下文信息, 提取出具有更强判别性的描述子。为了提取局部信息, 首先将多尺度特征 F_1 , F_2 和 F_3 下采样至 $H/4 \times W/4$; 然后将下采样后的多尺度特征沿着通道维度串接起来; 最后使用 1×1 卷积充分地融合多尺度局部信息, 得到特征图 D_s 。为了提取长距离上下文信息, 本文提出 LKCRF 模块, 其由 2 个高效的大卷积核分支 (efficient large kernel convolution21, 简称 ELKC21, 以及 efficient large kernel convolution9, 简称 ELKC9), 1 个 3×3 卷积分支, 以及 1 个残差连接分支组成; 每个大卷积核分支均由 2 个深度可分离卷积和 1 个 1×1 卷积组成。大卷积核分支能够感知更大范围的区域, 捕捉长距离依赖信息, 从而提取出对重复纹理等困难场景更加鲁棒的特征。ELKC21, ELKC9, 3×3 卷积分支和残差连接分支的输出张量之和即为 LKCRF 模块的输出量

$$D_l = \text{ELKC21}(F_3) \oplus \text{ELKC9}(F_3) \oplus \text{Conv3}(F_3) \oplus F_3 \quad (1)$$

其中, $\oplus$ 表示对应元素相加, $\text{Conv3}(\cdot)$ 表示 3×3 卷积层。

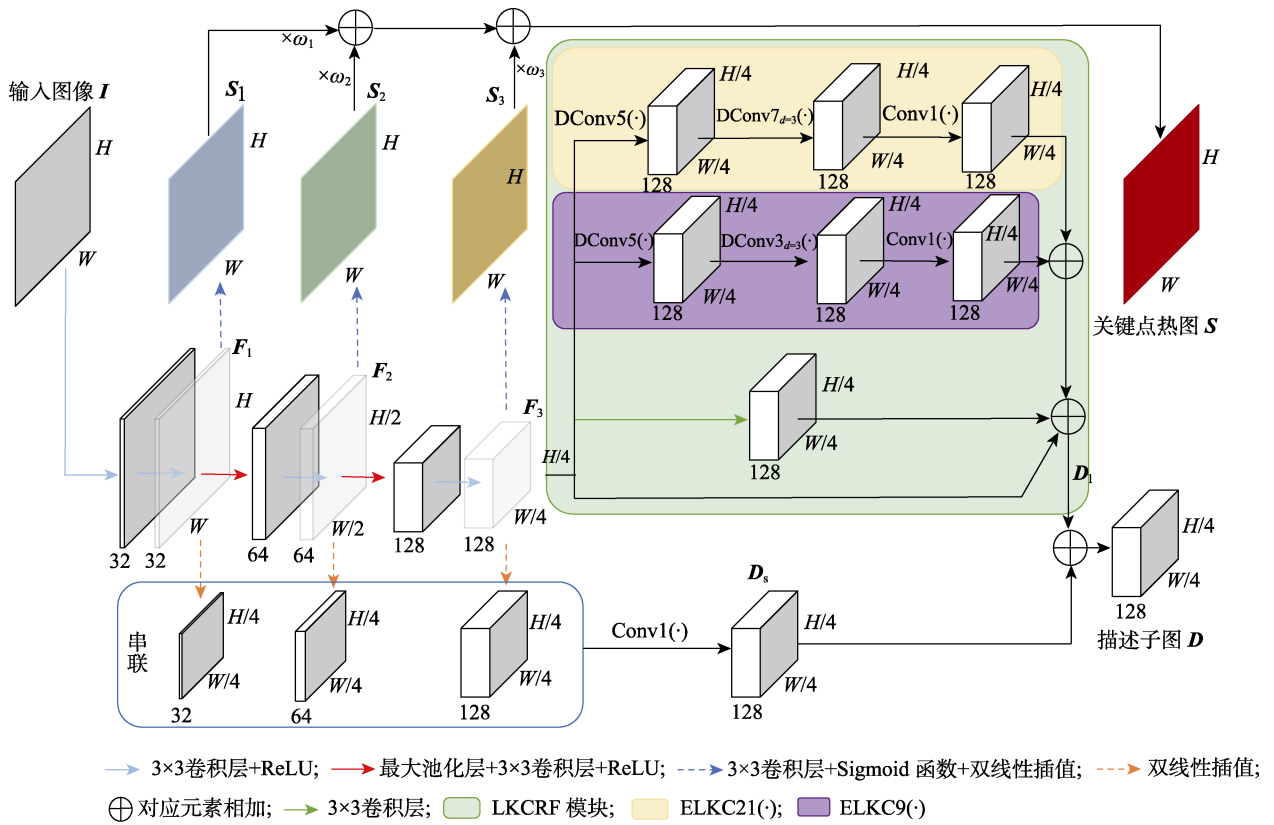


图 1 LKCRF-Net 的结构

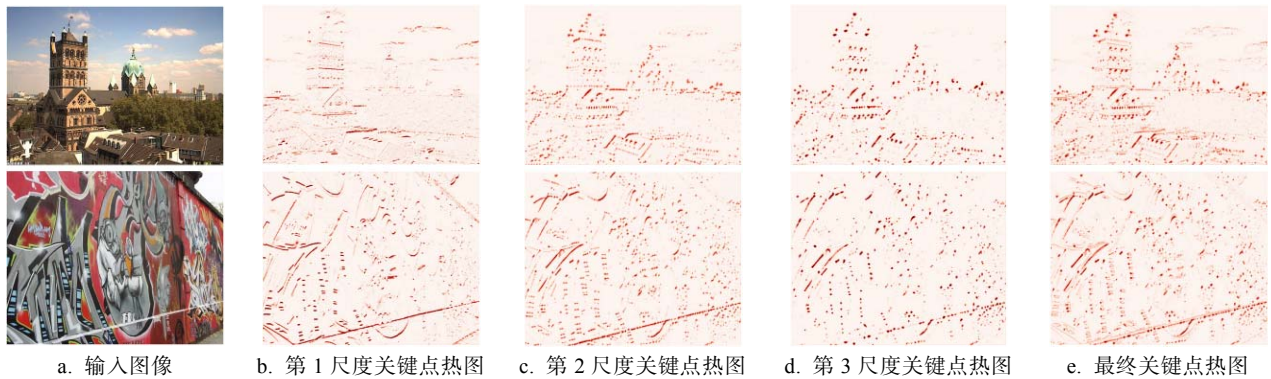


图 2 LKCRF-Net 输出的多尺度关键点热图

由于直接增大普通卷积操作的卷积核会大幅增大计算量和参数量, 因此, 本文采用高效大核卷积层^[21]构建 LKCRF 模块. 一个普通的 $K \times K$ 卷积被近似分解为 3 个部分: (1) 1 个 $K/d \times K/d$ 的深度可分离空洞卷积, 其中, d 表示扩张系数; (2) 1 个 $(2d-1) \times (2d-1)$ 的深度可分离卷积; (3) 1 个普通 1×1 卷积. 董荣胜等^[22]和魏文晓等^[23]也使用空洞卷积增大卷积神经网络的感受野. 与一个普通的 $K \times K$ 卷积相比, 上述近似分解过程大幅减小了计算复杂度和参数量. LKCRF 模块中, ELKC21 是一个普通 21×21 卷积的近似, 由 1 个 5×5 的深度可分

离卷积、1 个 7×7 的深度可分离空洞卷积(扩张系数为 3)和 1 个普通 1×1 卷积组成; ELKC9 是一个普通 9×9 卷积的近似, 由 1 个 5×5 的深度可分离卷积、1 个 3×3 的深度可分离空洞卷积(扩张系数为 3)和 1 个普通 1×1 卷积组成. ELKC21 和 ELKC9 的计算过程分别为

$$\text{ELKC21}(\cdot) = \text{Conv1}(\text{DConv7}_{d=3}(\text{DConv5}(\cdot))) \quad (2)$$

$$\text{ELKC9}(\cdot) = \text{Conv1}(\text{DConv3}_{d=3}(\text{DConv5}(\cdot))) \quad (3)$$

其中, $\text{Conv1}(\cdot)$ 表示普通 1×1 卷积; $\text{DConv7}_{d=3}(\cdot)$ 表示扩张系数为 3 的 7×7 深度可分离空洞卷积;

DConv3_{d=3}(·)表示扩张系数为 3 的 3×3 深度可分离空洞卷积; DConv5(·)表示 5×5 的深度可分离卷积。

2.2 全图像范围内自适应最困难描述子三元组构建方法

基于学习的图像特征点提取方法利用三元组损失函数训练神经网络的描述子提取能力; 而 ASLFeat^[13]仅在一对训练图像的共视区域内部通过随机选择匹配点对构建描述子三元组, 其训练设置与实际应用中的特征点匹配任务有很大差异。在实际应用中, 由于共视区域通常是未知的, 特征点匹配任务要求描述子具有在整幅图像范围内确定正确匹配点的能力, 而不仅仅在共视区域内寻找正确匹配点。因此, 为了进一步提升描述子的匹配性能, 本文提出一种全图像范围内自适应最困难描述子三元组构建(image adaptive hardest triplet, IAHT)方法。

给定一对训练图像 I_1 和 I_2 , LKCRF-Net 估计出这对训练图像的关键点热图 $S^{(1)}$ 和 $S^{(2)}$, 以及描述子图 D_1 和 D_2 。在实际应用中, 基于学习的图像特征点提取方法采用非极大值抑制方法检测关键点。为了在训练过程中模拟非极大值抑制方法, 并针对关键点优化描述子, 本文将 $S^{(1)}$ 和 $S^{(2)}$ 均匀划分为互不重叠的网格单元, 网格单元尺寸为 $N \times N$ 。以 $S^{(1)}$ 为例, 在每个网格单元中, 用 Soft-argmax 操作提取一个关键点。给定一个 $N \times N$ 的网格单元, 其内部提取关键点的计算公式为

$$s'(i, j) = \frac{\exp(T_{\text{det}} s(i, j))}{\sum_{i'=1}^N \sum_{j'=1}^N \exp(T_{\text{det}} s(i', j'))} \quad (4)$$

$$[\hat{i}, \hat{j}]^T = \sum_{0 \leq i, j \leq N} s'(i, j) [i, j]^T \quad (5)$$

其中, $s(i, j)$ 表示 $S^{(1)}$ 在位置 (i, j) 的取值; $[\hat{i}, \hat{j}]^T$ 即从该网格单元中提取的关键点坐标; $T_{\text{det}}=50$ 表示温度系数。

根据已知的 I_1 和 I_2 之间的稠密匹配关系, 可以在 I_2 中确定 $S^{(1)}$ 中的关键点的正确对应点。以 $S^{(1)}$ 中的某个关键点 P_{li} 为参考点, 其在 I_2 中的正确对应点为 $P_{li}^{(2)}$ 。 $\{P_{2j}\}_j^M$ 表示从 I_2 中提取的关键点集合, 值得注意的是, $\{P_{2j}\}_j^M$ 分布于 I_2 的所有区域, 而不仅仅处于 I_2 的共视区域中。 d_{li} 表示 P_{li} 的描述子, $d_{li}^{(2)}$ 表示 $P_{li}^{(2)}$ 的描述子, d_{2j} 表示 P_{2j} 的

描述子。正确匹配点对的描述子欧几里得距离为

$$p_{li} = \|d_{li} - d_{li}^{(2)}\|_2 \quad (6)$$

最困难非匹配点对的描述子距离为

$$n_{li} = \min(\|d_{li} - d_{2E}\|_2, \|d_{li}^{(2)} - d_{2E}\|_2) \quad (7)$$

$$\begin{cases} E = \arg \min_{j \in [1, M]} (\min(\|d_{li} - d_{2j}\|_2, \|d_{li}^{(2)} - d_{2j}\|_2)) \\ \text{s.t. } \|P_{2j} - P_{li}^{(2)}\|_2 > R \end{cases} \quad (8)$$

其中, $R=12$ 表示安全半径^[13]。上述方法确定了描述子三元组 $(d_{li}, d_{li}^{(2)}, d_{2E})$ 。现有方法仅在 2 幅训练图像之间选择非匹配点对, 即只将 d_{li} 当作锚描述子; 与之相比, 本文方法同时考虑图像对之间和单幅图像内部的非匹配点对, 能够自适应地选择 d_{li} 或 $d_{li}^{(2)}$ 为最佳的锚描述子, 构建出更加困难的描述子三元组用于训练, 从而能够产生具有更强判别性的描述子。该三元组的损失函数项为

$$L_{\text{des}, li} = [p_{li} - m_p]_+ + [m_n - n_{li}]_+ \quad (9)$$

其中, $[x]_+ = \max(0, x)$; $m_p = 0.2$ 表示正确匹配描述子的欧几里得距离裕量; $m_n = 1.0$ 表示错误匹配描述子的欧几里得距离裕量。 $S^{(1)}$ 中的所有共视关键点均可被选定为参考点, 本文为 $S^{(1)}$ 中的所有参考点计算如式(9)的损失项。

$$L_{\text{des1}} = \frac{1}{C} L_{\text{des}, li} \quad (10)$$

表示 $S^{(1)}$ 中的所有参考点的损失项的平均值。其中, C 表示 $S^{(1)}$ 中参考点的个数。类似地, 本文还选定 $S^{(2)}$ 中的所有共视关键点为参考点, 计算出 L_{des2} 。最终的描述子损失函数为

$$L_{\text{des}} = \frac{1}{2} (L_{\text{des1}} + L_{\text{des2}}) \quad (11)$$

2.3 用于学习关键点检测的损失函数

依据第 2.2 节, 本文将 $S^{(1)}$ 和 $S^{(2)}$ 分为网格, 并利用 Soft-argmax 函数从每个网格中提取一个关键点。以 $S^{(1)}$ 中的某个共视关键点 P_{li} 为参考点, 依据已知的 I_1 和 I_2 之间的稠密匹配关系, 可以得到 P_{li} 在 I_2 中的对应点 $P_{li}^{(2)}$ 。假设 P_{2i} 是 I_2 中与 $P_{li}^{(2)}$ 距离最近的关键点, 且 $\|P_{2i} - P_{li}^{(2)}\|_2 \leq t_{kp}$, t_{kp} 表示距离阈值, $t_{kp}=8$ 。 $P_{li}^{(2)}$ 与 P_{2i} 的欧几里得距离即以 P_{li} 为参考点的关键点学习损失项, 公式为

$$L_{kp, li} = \|P_{2i} - P_{li}^{(2)}\|_2 \quad (12)$$

$S^{(1)}$ 中的所有共视关键点均可被当作参考点, 它们的损失函数项平均值公式为

$$L_{kp1} = \frac{1}{K} L_{kp,li} \quad (13)$$

其中, K 表示 $S^{(1)}$ 中的共视关键点的个数. 类似地, 本文还将 $S^{(2)}$ 中的所有共视关键点当作参考点, 得到 L_{kp2} . 最终的用于学习关键点检测的损失函数公式为

$$L_{kp} = \frac{1}{2} (L_{kp1} + L_{kp2}) \quad (14)$$

总体的损失函数公式为

$$L = \lambda_{des} L_{des} + \lambda_{kp} L_{kp} \quad (15)$$

其中, λ_{kp} 和 λ_{des} 都表示权重, $\lambda_{kp} = 10$, $\lambda_{des} = 1$.

3 实验及结果分析

3.1 实现细节

仿照 D2-Net^[12], 本文使用 MegaDepth 数据集^[24]训练 LKCRF-Net. MegaDepth 数据集中包含图像的相机位姿、相机内参数和深度图, 由此, 可以得到训练图像对之间的稠密匹配关系. 仿照 D2-Net, 本文使用 117 个场景进行训练, 使用 77 个场景进行验证, 用于训练和验证的场景与 D2-Net 相同. 每个场景中, 随机选择 100 对图像作为训练图像对, 训练图像的尺寸为 560×560 像素. 本文使用 AdamW 算法训练网络, 初始的学习率为 0.0003, 训练回合数为 70, 批处理大小为 4, 网格单元边长 $N = 8$.

3.2 测试数据集

(1) HPatches 数据集^[1]. 广泛用于评测图像特征点提取方法的匹配性能和单应矩阵估计精度, 其含有图像对之间的真值单应矩阵. 仿照 D2-Net, 本文选取其中的 108 个图像序列进行评测, 其中, 56 个图像序列包含显著视点变化, 52 个图像序列包含剧烈光照变化.

(2) Aachen Day-Night v1.1 数据集^[3]. 广泛用于评测图像特征点提取方法在室外大场景视觉定位任务中的表现, 其包含 6697 幅参考图像和 191 幅待定位的夜晚图像; 其中的视觉定位任务要求图像特征点提取方法估计白天图像和夜晚图像之间匹配点对, 能够充分地评测图像特征点提取方法对于剧烈光照变化情况的鲁棒性.

(3) InLoc 数据集^[25]. 广泛用于评测图像特征点提取方法在室内场景视觉定位任务中的表现,

其包含 356 幅由 iPhone7 手机拍摄的待定位图像和 9972 幅参考图像; 含有剧烈的视点变化、光照变化、重复纹理、弱纹理等具有挑战性的场景.

3.3 图像匹配和单应矩阵估计

利用 HPatches 数据集^[1]评测本文方法在图像匹配和单应矩阵估计任务中的表现. 采用已有研究工作常用的 3 个标准评测指标: (1) 匹配得分 (matching score, MS). 表示在一定误差容错范围内, 共视区域中的正确匹配数与关键点数的比值. (2) 平均匹配准确率 (mean matching accuracy, MMA). 表示在一定误差容错范围内, 正确匹配数与双向最近邻方法估计出的匹配数的比值. (3) 单应矩阵准确率 (homography accuracy, HA). 表示在一定误差容错范围内, 估计出的正确单应矩阵数与单应矩阵总数的比值. 给定一幅图像, 每种图像特征点提取方法至多提取 5000 个特征点.

(1) 通过消融实验, 证明 LKCRF 模块和 IAHT 的有效性, 其中, 不同模型的设置情况如表 1 所示. 将图 1 中的 LKCRF 模块替换为 3 个连续的普通 3×3 卷积+ReLU, 得到模型 1; 模型 2 的网络结构与 LKCRF-Net 相同, 仅在共视区域内部构建描述子三元组 HT_cov; 模型 3 的网络结构与 LKCRF-Net 相同, 但是仅采用一对训练图像之间的非匹配点对 (image hardest triplet, IHT), 即将 d_{li} 当作锚描述子; 将 LKCRF-Net 中的 LKCRF 模块替换为 Inception 模块^[26], 得到模型 4; 还将不同尺度的权重 ω_1 , ω_2 和 ω_3 设置为可学习参数, 得到模型 5. LKCRF-Net 采用本文提出的 IAHT 作为描述子损失函数. 在消融实验中, 所有模型的关键点损失函数均为 L_{kp} .

表 1 消融实验中不同模型的设置

模型	网络结构	描述子和关键点损失函数	
		描述子	关键点
1	无 LKCRF 模块	IAHT	L_{kp}
2	LKCRF-Net	HT_cov	L_{kp}
3	LKCRF-Net	IHT	L_{kp}
4	Inception	IAHT	L_{kp}
5	权重 ω 可学习	IAHT	L_{kp}
LKCRF-Net	LKCRF-Net	IAHT	L_{kp}

表 1 中的 6 个模型在 HPatches 数据集上的消融实验结果如表 2 所示, 其中, 容错阈值为 3 个像素. 比较模型 1 和 LKCRF-Net 可得, LKCRF 模块大幅提升了网络的图像匹配精度和单应矩阵估计

的准确率, 证明 LKCRF 模块能够利用长距离上下文信息增强局部描述子的判别性. 值得注意的是, 模型 1 的参数量为 613.00K, 当输入一幅尺寸为 640×480 像素的图像时, 其浮点数运算次数为 17.96G; 虽然 LKCRF-Net 含有大卷积核, 但是其参数量仅为 382.98K, 当输入一幅尺寸为 640×480 像素的图像时, 其浮点数运算次数仅为 13.53G. 因此, 在参数量和计算复杂度远低于模型 1 的情况下, LKCRF-Net 取得了更加优越的性能. 虽然 LKCRF-Net 的 MS 得分低于模型 2 和模型 3, 但是其 MMA 和 HA 得分均高于模型 2 和模型 3. LKCRF-Net 取得优越性能的原因如下: 与 HT_cov 相比, IAHT 将描述子三元组构建范围由共视区域扩展至整幅图像范围, 使描述子训练过程与实际应用场景更加相似; 与 IHT 相比, IAHT 能自适应地选择最佳的锚描述子, 构建出更有价值的描述子三元组, 提升描述子的判别性; LKCRF-Net 的实验结果显著优于模型 4, 充分地表明与 Inception 结构相比, LKCRF 模块有助于网络捕捉不同尺寸大小的图像区域信息, 生成具有更强判别性的描述子; 与模型 5 相比, LKCRF-Net(手动设定权重)取得了更高的匹配精度和单应矩阵估计精度, 这是因为将不同尺度的权重 ω_1 , ω_2 和 ω_3 设置为可学习参数造成网络过拟合, 训练结束后, $\omega_1 = 5.15$,

$\omega_2 = 0.56$, $\omega_3 = 0.27$, 网络过于重视第 1 尺度的信息, 而忽略了另外 2 个尺度上的信息, 导致模型性能下降.

表 2 在 HPatches 数据集上的消融实验结果 %

模型	MMA	MS	HA
1	70.39	40.04	72.04
2	74.76	42.31	70.56
3	73.67	43.47	72.41
4	66.56	35.18	71.67
5	70.71	40.76	71.48
LKCRF-Net	75.15	40.42	76.11

注: 粗体表示最优值.

(2) 将 LKCRF-Net 与其他图像特征点提取方法进行比较, 包括 RootSIFT^[5,27], SuperPoint^[11], D2-Net^[12], ASLFeat^[13], R2D2^[14]和 ALIKE^[15]. 一些定性实验结果如图 3 所示, 其中, 绿色线段表示符合单应矩阵真值的正确匹配, $\{\cdot\}$ 表示正确匹配数量. 可以看出, 在重复纹理场景中, LKCRF-Net 估计出的正确匹配数量显著大于现有的基于卷积神经网络的图像特征点提取方法, 原因是与基于卷积神经网络的方法相比, LKCRF-Net 的描述子输出端能够感受更大范围的图像区域, 融合长距离上下文信息和局部信息, 提取出对重复纹理场景更加

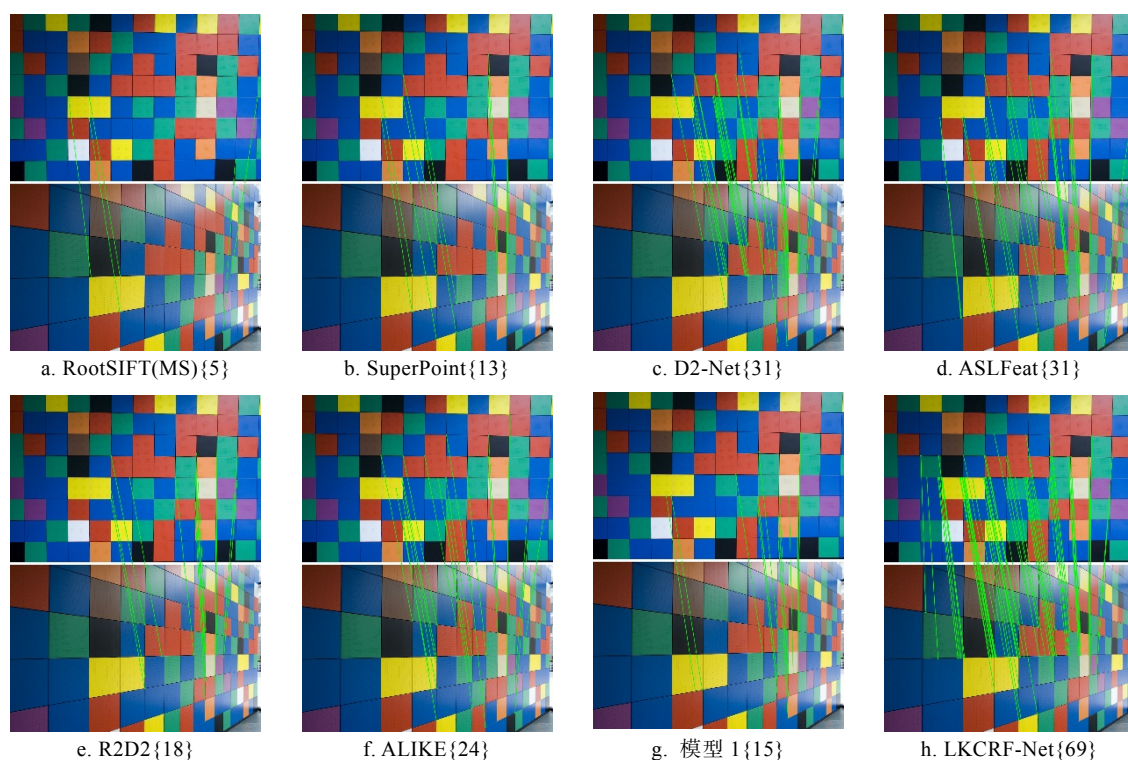


图 3 8 种图像特征点提取算法在 HPatches 数据集上的定性实验结果

鲁棒的描述子. 定量结果如表 3 所示, 其中, MS 表示从多个尺度上提取特征点. 可以看出, 在容误差为 3 像素和 5 像素时, LKCRF-Net 取得了最高的 MMA 和 HA 得分, 即取得了最佳的匹配精度和单应矩阵估计准确率.

表 3 7 种方法在 HPatches 数据集上的实验结果

方法	MMA			MS			HA		
	1 像素	3 像素	5 像素	1 像素	3 像素	5 像素	1 像素	3 像素	5 像素
RootSIFT(MS) ^[5,27]	35.35	51.37	54.77	17.85	25.38	26.95	47.59	74.26	83.70
SuperPoint ^[11]	34.30	65.39	73.80	23.06	42.93	48.14	35.56	71.85	83.70
D2-Net ^[12]	11.74	43.58	66.66	6.26	21.93	32.74	6.48	35.74	62.96
ASLFeat ^[13]	39.79	72.26	82.13	24.19	41.95	46.98	37.96	72.59	83.89
R2D2 ^[14]	29.88	66.92	74.47	13.18	26.87	29.34	32.04	64.26	75.74
ALIKE ^[15]	41.50	68.10	74.23	24.03	37.90	40.98	42.96	75.18	84.26
LKCRF-Net	40.09	75.15	82.27	22.59	40.42	43.75	40.56	76.11	85.37

注: 粗体表示最优值.

本文采用 Aachen Day-Night v1.1 数据集官方提供的流程代码估计图像的相机位姿, 首先导入图像特征点数据, 然后利用 COLMAP^[2]进行图像配准; 采用在(0.25 m, 2°), (0.5 m, 5°)和(5 m, 10°)这 3 组误差阈值范围内的图像定位成功率作为评测指标. 仿照现有的工作, 给定一幅图像, 每种图像特征点提取方法至多提取 20 000 个特征点. 定位结果如表 4 所示. 可以看出, 在较严格的误差阈值范围(0.25 m, 2°)和(0.5 m, 5°)内, LKCRF-Net 取得了最高的定位成功率; 在(5 m, 10°)误差阈值范围内, LKCRF-Net 仅次于 ALIKE. 实验结果表明, LKCRF-Net 对剧烈光照变化和视角变化情况鲁棒.

表 4 7 种方法在 Aachen Day-Night v1.1 数据集上的定位结果

方法	描述子 维数	误差阈值范围		
		(0.25 m, 2°)	(0.5 m, 5°)	(5 m, 10°)
RootSIFT ^[5,27]	128	53.4	62.3	72.3
SuperPoint ^[11]	256	68.1	85.9	94.8
D2-Net ^[12]	512	67.0	86.4	97.4
ASLFeat ^[13]	128	71.2	85.9	96.9
R2D2 ^[14]	128	70.7	85.3	96.9
ALIKE ^[15]	128	67.5	85.3	97.9
LKCRF-Net	128	73.3	86.9	97.4

注: 粗体表示最优值.

3.5 室内大规模视觉定位

采用 InLoc 数据集^[25]评测本文方法在室内视觉定位任务中的性能. 将 LKCRF-Net 与在 HPatches 数据集上相同的对比方法进行比较.

3.4 室外大场景视觉定位

采用 Aachen Day-Night v1.1 数据集^[3]评测本文方法在室外大场景视觉定位任务中的性能. 将 LKCRF-Net 与在 HPatches 数据集上相同的对比方法进行比较.

将图像特征点数据导入分层定位工具箱^[28]中进行图像位姿估计; 采用在(0.25 m, 10°), (0.5 m, 10°)和(1 m, 10°)这 3 组误差阈值范围内的图像定位成功率作为评测指标. 值得注意的是, InLoc 数据集中含有很多重复纹理和弱纹理图像. 在 InLoc 数据集中的特征点匹配可视化结果如图 4 所示. 可以看出, 在重复纹理和弱纹理区域, LKCRF-Net 估计出的正确匹配数大于 ALIKE^[15]估计出的正确匹配数. 定位定量结果如表 5 所示, 其中, DUC1 和 DUC2 表示 InLoc 数据集中的 2 个不同场景. 可以看出, LKCRF-Net 取得了最高的室内视觉定位精度. 图 4 和表 5 验证了 LKCRF-Net 对于重复纹理和弱纹理场景的鲁棒性.

3.6 石油生产场景中的评测实验

利用自采的石油生产场景数据评测 LKCRF-Net 的性能. 针对室外油田场景图像含有很多空旷的弱纹理区域, 室内生产场景图像含有很多管道等重复纹理区域的特点, 本文从室外油田场景和室内生产场景中分别随机选择 100 对含有共视区域的图像. 由于自采数据缺乏图像对稠密匹配真值, 因此本文遵循 Fan 等^[18]的评测方法, 比较经 RANSAC 筛选后的、不同图像特征点提取方法的特征点匹配内点数量. 为了保证公平比较, 每种特征点提取方法从 1 幅图像中至多提取 5 000 个特征点, 并采用最近邻匹配方法估计特征点匹配. 定量实验结果如表 6 所示. 可以看出, 在 2 种场景中, LKCRF-Net 均取得了最大的平均特征点匹配内点数, 证明了其在石油生产场景中的优势.

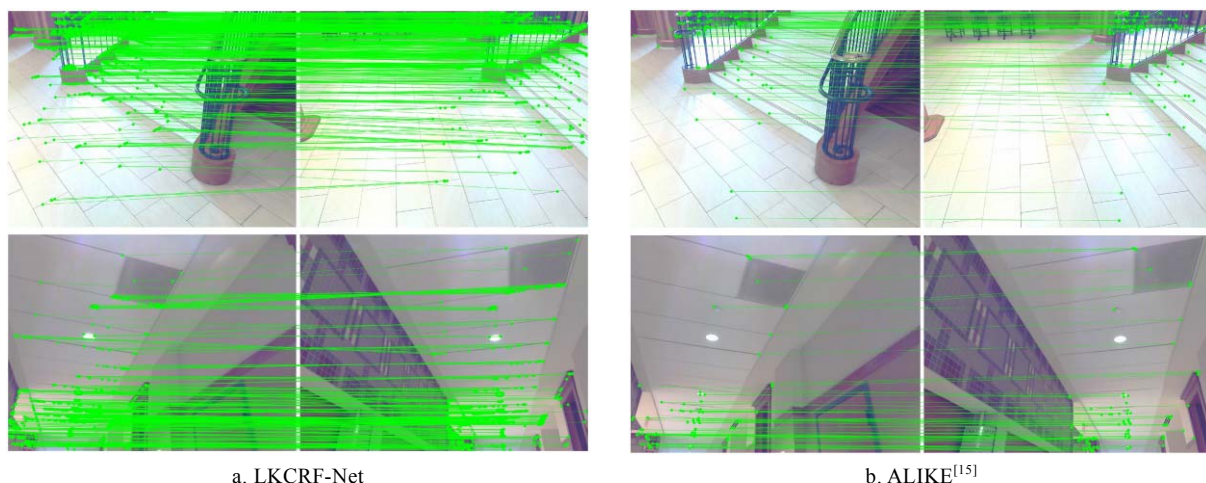


图 4 室内大规模视觉定位任务中的特征点匹配可视化结果图

表 5 7 种方法在 InLoc 数据集上的定位结果

%

方法	DUC1			DUC2		
	(0.25 m, 10°)	(0.5 m, 10°)	(1 m, 10°)	(0.25 m, 10°)	(0.5 m, 10°)	(1 m, 10°)
RootSIFT ^[5,27]	14.1	23.7	27.8	3.8	6.9	8.4
SuperPoint ^[11]	36.9	58.1	68.7	34.4	50.4	61.8
D2-Net ^[12]	29.8	52.5	66.7	35.1	48.9	57.3
ASLFeat ^[13]	36.4	56.1	66.7	36.6	55.7	61.1
R2D2 ^[14]	32.3	52.5	64.6	29.8	41.2	48.9
ALIKE ^[15]	29.3	46.5	59.6	25.2	38.9	47.3
LKCRF-Net	38.9	56.1	69.2	40.5	55.7	64.1

注: 粗体表示最优值。

表 6 在石油生产场景中的平均特征点匹配内点数

方法	室外油田场景	室内生产场景	方法	室外油田场景	室内生产场景
SuperPoint ^[11]	92.34	284.07	R2D2 ^[14]	111.90	390.70
D2-Net ^[12]	106.98	443.07	ALIKE ^[15]	162.64	270.29
ASLFeat ^[13]	145.86	549.07	LKCRF-Net	195.32	655.06

注: 粗体表示最优值。

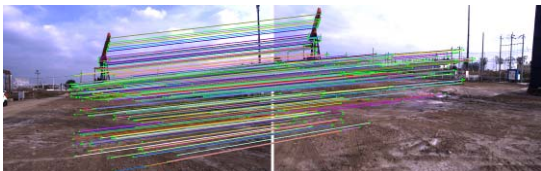
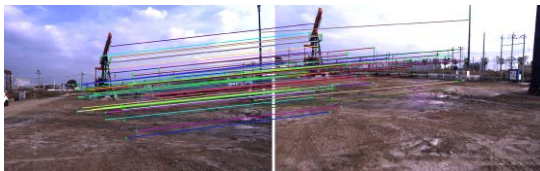
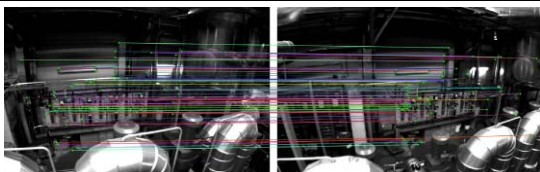
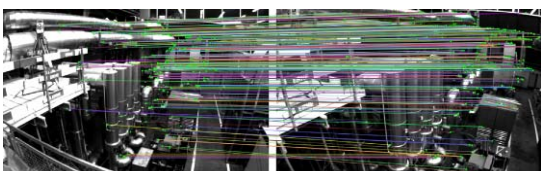
在石油生产场景中的定性实验结果如表 7 所示。可以看出, 在含有弱纹理区域的室外空旷油田场景和含有重复纹理区域的室内生产场景中, LKCRF-Net 估计出的特征点的正确匹配数均大于 ALIKE^[15]估计出的正确匹配数, 证明了 LKCRF-Net 在石油生产场景的优越性。

4 结 语

本文提出一种融合局部和长距离信息的图像特征点提取方法。首先提出一个 LKCRF 模块, 并利用该模块构建了一个高效的图像特征点提取网络。相较于文中对比的基于卷积神经网络的图像特征点提取方法, 该模块能够感知更大范围的图

像区域, 充分地利用长距离上下文信息来增强图像局部描述子的判别性, 估计出高精度的图像特征点匹配。然后提出一种 IAHT 方法, 不仅可以描述子三元组的构建范围由共视区域扩展至整幅图像, 而且能自适应地选择最难以区分的非匹配点对, 构建出更有训练价值的描述子三元组, 进一步提升图像特征点的匹配性能。在图像匹配、单应矩阵估计、室外大场景视觉定位和室内大规模视觉定位任务中的实验结果表明, LKCRF-Net 优于对比方法。但是, 与所有基于学习的图像特征点提取方法相同, LKCRF-Net 在部分算力较低的终端设备上尚不能实时运行。未来, 将研究更加轻量化的图像特征点提取方法和性能更优越的多尺度特征融合方法。

表 7 在石油生产场景图像中的匹配可视化结果

场景	LKCRF-Net	ALIKE ^[15]
室外		
		
室内		
		

参考文献(References):

- [1] Balntas V, Lenc K, Vedaldi A, *et al.* HPatches: a benchmark and evaluation of handcrafted and learned local descriptors[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3852-3861
- [2] Schönberger J L, Frahm J M. Structure-from-motion revisited[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4104-4113
- [3] Zhang Z C, Sattler T, Scaramuzza D. Reference pose generation for long-term visual localization via learned features and view synthesis[J]. *International Journal of Computer Vision*, 2021, 129(4): 821-844
- [4] Campos C, Elvira R, Rodríguez J J G, *et al.* ORB-SLAM3: an accurate open-source library for visual, visual-inertial, and multimap SLAM[J]. *IEEE Transactions on Robotics*, 2021, 37(6): 1874-1890
- [5] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110
- [6] Harris C, Stephens M. A combined corner and edge detector[OL]. [2023-11-19]. <https://web.stanford.edu/class/cs231m/references/harris-stephens.pdf>
- [7] Rosten E, Drummond T. Machine learning for high-speed corner detection[C] //Proceedings of the 9th European Conference on Computer Vision. Heidelberg: Springer, 2006: 430-443
- [8] Alcantarilla P F, Bartoli A, Davison A J. KAZE features[C] //Proceedings of the 12th European Conference on Computer Vision. Heidelberg: Springer, 2012: 214-227
- [9] Rublee E, Rabaud V, Konolige K, *et al.* ORB: an efficient alternative to SIFT or SURF[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 2564-2571
- [10] Calonder M, Lepetit V, Strecha C, *et al.* BRIEF: binary robust independent elementary features[C] //Proceedings of the 11th European Conference on Computer Vision. Heidelberg: Springer, 2010: 778-792
- [11] DeTone D, Malisiewicz T, Rabinovich A. SuperPoint: self-supervised interest point detection and description[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2018: 337-349
- [12] Dusmanu M, Rocco I, Pajdla T, *et al.* D2-Net: a trainable CNN for joint description and detection of local features[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 8084-8093
- [13] Luo Z X, Zhou L, Bai X Y, *et al.* ASLFeat: learning local features of accurate shape and localization[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 6588-6597
- [14] Revaud J, Weinzaepfel P, De Souza C, *et al.* R2D2: repeatable and Reliable detector and descriptor[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: Article No.1113

- [15] Zhao X M, Wu X M, Miao J Y, *et al.* ALIKE: accurate and lightweight keypoint detection and descriptor extraction[J]. *IEEE Transactions on Multimedia*, 2023, 25: 3101-3112
- [16] Dusmanu M, Miksik O, Schönberger J L, *et al.* Cross-descriptor visual localization and mapping[C] // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. Los Alamitos: IEEE Computer Society Press, 2021: 6038-6047
- [17] Shen X L, Wang C, Li X, *et al.* Learning scale awareness in keypoint extraction and description[J]. *Pattern Recognition*, 2022, 121: Article No.108221
- [18] Fan B, Yang Y Z, Feng W S, *et al.* Seeing through darkness: visual localization at night via weakly supervised learning of domain invariant features[J]. *IEEE Transactions on Multimedia*, 2023, 25: 1713-1726
- [19] Wang Z C, Wu C X, Yang Y F, *et al.* Learning transformation-predictive representations for detection and description of local features[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2023: 11464-11473
- [20] Wang X J, Liu Z Y, Hu Y, *et al.* FeatureBooster: boosting feature descriptors with a lightweight neural network[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2023: 7630-7639
- [21] Guo M H, Lu C Z, Liu Z N, *et al.* Visual attention network[J]. *Computational Visual Media*, 2023, 9(4): 733-752
- [22] Dong Rongsheng, Liu Yi, Ma Yuqi, *et al.* Lightweight network with convolutional attention feature fusion for real-time semantic segmentation[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2023, 35(6): 935-943(in Chinese)
(董荣胜, 刘意, 马雨琪, 等. 轻量化卷积注意力特征融合网络的实时语义分割[J]. *计算机辅助设计与图形学学报*, 2023, 35(6): 935-943)
- [23] Wei Wenxiao, Liu Jieyu, Xu Junhui, *et al.* Small target detection algorithm based on receptive field amplification feature fusion[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2023, 35(1): 48-54(in Chinese)
(魏文晓, 刘洁瑜, 徐军辉, 等. 扩增感受野特征融合的小目标检测算法[J]. *计算机辅助设计与图形学学报*, 2023, 35(1): 48-54)
- [24] Li Z Q, Snavely N. MegaDepth: learning single-view depth prediction from internet photos[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2018: 2041-2050
- [25] Taira H, Okutomi M, Sattler T, *et al.* InLoc: indoor visual localization with dense matching and view synthesis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(4): 1293-1307
- [26] Szegedy C, Liu W, Jia Y Q, *et al.* Going deeper with convolutions[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2015: 1-9
- [27] Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2012: 2911-2918
- [28] Sarlin P E, Cadena C, Siegwart R, *et al.* From coarse to fine: robust hierarchical localization at large scale[C] // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Los Alamitos: IEEE Computer Society Press, 2019: 12708-12717