基于重复性和特异性约束的图像特征匹配

郭印宏，王立春，李爽

(北京工业大学信息学部，北京 100124)

摘　　要：图像特征匹配通过比较一对像素在特征空间的距离确定其是否可匹配，如何学习鲁棒的像素特征是基于深度学习的图像特征匹配要解决的关键问题之一，另外，像素特征表示的学习也受到源图像质量的影响。针对学习更鲁棒的像素特征表示的问题，对图像特征匹配网络 LoFTR 进行改进。针对粗粒度特征重构分支，定义特异性约束使得同一幅图像内像素的特征距离尽可能远，使不同像素间具有强区分性；定义重复性约束使得不同图像的匹配点对的特征距离尽可能近，使不同图像间的匹配像素点具有强相似性，以增强匹配的准确性。在 Backbone 的解码阶段增加图像重建层，定义图像重建损失约束编码器学习更鲁棒的特征表示。在室内数据集 ScanNet 与室外数据集 MegaDepth 上的实验结果证明了本文方法的有效性，构建了不同质量图像数据并验证了方法能够更好地适应不同质量图像的特征匹配。

关　键　词：深度学习；图像特征匹配；重复性；特异性；图像重建损失

中图分类号：TP 391　　**DOI**：10.11996/JG.j.2095-302X.2023040739

文献标识码：A　　**文 章 编 号**：2095-302X(2023)04-0739-08

Image feature matching based on repeatability and specificity constraints

GUO Yin-hong, WANG Li-chun, LI Shuang

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Image feature matching ascertains whether a pair of pixels can be matched by comparing their distance in the feature space. Therefore, how to learn robust pixel features constitutes one of the primary concerns in the field of image feature matching based on deep learning. In addition, the learning of pixel feature representation is also affected by the quality of the source image. As a solution to the challenge of learning more robust pixel feature representations, the proposed method improved the image feature matching network LoFTR. For the coarse granularity feature reconstruction branch, the specificity constraint was defined to maximize the feature distance between pixels within the same image, enabling strong distinguishability between different pixels. The repeatability constraint was defined to minimize the feature distance between the matched pixels from different images, enabling strong similarity between the matched pixels across different images and thus enhancing the accuracy of matching. Additionally, an image reconstruction layer was incorporated into the decoding phase of the Backbone, and image reconstruction loss was

收稿日期：2022-11-28；**定稿日期**：2023-04-06

Received：28 November，2022；**Finalized**：6 April，2023

基金项目：科技创新2030-“新一代人工智能”重大项目(2021ZD0111902)；国家自然科学基金项目(U21B2038，61876012，62172022)；中国高校产学研创新基金项目(2021JQR023)

Foundation items：Science and Technology Innovation 2030 – “New Generation of Artificial Intelligence” Major Project (2021ZD0111902); National Natural Science Foundation of China (U21B2038, 61876012, 62172022); Foundation for China University Industry-University Research Innovation (2021JQR023)

第一作者：郭印宏(1997–)，男，硕士研究生。主要研究方向为计算机视觉。E-mail：gyh20200216@163.com

First author：GUO Yin-hong (1997–), master student. His main research interest covers computer vision. E-mail：gyh20200216@163.com

通信作者：王立春(1975–)，女，教授，博士。主要研究方向为计算机视觉、人机交互等。E-mail：wanglc@bjut.edu.cn

Corresponding author：Wang Li-chun (1975–), professo, Ph.D. Her main research interests cover computer vision and human-computer interaction, etc. E-mail：wanglc@bjut.edu.cn

defined to constrain the encoder to learn more robust feature representation. The experimental results on indoor dataset ScanNet and outdoor dataset MegeDepth show the effectiveness of the proposed method. Furthermore, based on images with different qualities, it is verified that the proposed method can better adapt to image feature matching when the source images have different quality.

Keywords: deep learning; image feature matching; repeatability; specificity; image reconstruction loss

图像特征匹配是许多 3D 计算机视觉任务的基础，如同步定位与地图构建(simultaneous localization and mapping，SLAM)[1]、视觉定位等。给定一对要匹配的图像，大多数现有的匹配方法包括特征检测、特征描述和特征匹配 3 个独立的阶段。特征检测阶段将图像中的角点作为关键点；特征描述阶段基于关键点的邻域提取局部描述子，一对图像在特征检测和描述阶段产生 2 组关键点及对应的描述子；特征匹配阶段通常利用最近邻搜索或更复杂的匹配算法计算 2 幅图像中点和点的匹配关系。特征检测器的使用减小了特征匹配的搜索空间，且得到的稀疏匹配可以满足一些任务(如相机位姿估计)的需求。但由于环境因素的影响，在图像纹理较弱、图像中存在重复的物体、视点或光照变化较大、运动模糊等情况下，基于检测器的图像特征匹配方法可能无法提供足够多的关键点，从而对位姿估计、视觉定位等任务产生不利影响。

现有的基于深度学习的无检测器图像特征匹配方法在图像弱纹理区域、视点变化等情况下可以得到质量较好的匹配点对，如 LoFTR[2]方法。然而，当图像中弱纹理区域较多时，该方法的性能会有所降低，尤其对于模糊图像其性能下降明显。为此，本文提出了一种基于重复性和特异性约束的图像特征匹配方法，可以更好地适应不同质量图像的特征匹配任务，无论是弱纹理区域较多的图像，还是模糊图像，都有较好性能。

本文实现了一个无检测器的图像特征匹配网络，采用由粗到细的分层匹配策略，如图 1 所示。具体步骤如下：

步骤 1. 利用 Backbone 提取图像特征，并进行图像重建。

步骤 2. 步骤 1 提取的低分辨率(输入图像尺寸的 1/8)特征($\boldsymbol{F}_1^A,\boldsymbol{F}_1^B$)输入到粗粒度特征重构模块中进行特征重构，约束重构后的特征($\boldsymbol{F}_{tr1}^A,\boldsymbol{F}_{tr1}^B$)具有重复性和特异性。重复性指 2 幅图像中匹配点的特征之间相似度较高，约束匹配的一对点之间具有相似性。特异性指同一幅图像中不同像素特征之间的差异较大，约束不同像素之间具有较强的区分度。重构特征输入到可微匹配层计算得到粗粒度匹配点对集 $\boldsymbol{M}_C=\{(\tilde{i},\tilde{j})\}$。

步骤 3. 将步骤 2 得到的匹配点对映射到高分辨率(输入图像尺寸的 1/2)的特征图中，得到位置 $\hat{i}$ 和 $\hat{j}$。以 $\hat{i}$ 和 $\hat{j}$ 为中心，将一定范围内的特征输入到细粒度特征重构模块进行特征重构，基于重构后的特征($\boldsymbol{F}_{tr2}^A,\boldsymbol{F}_{tr2}^B$)计算得到精细化的匹配点对集 $\boldsymbol{M}=\{(i,j)\}$。

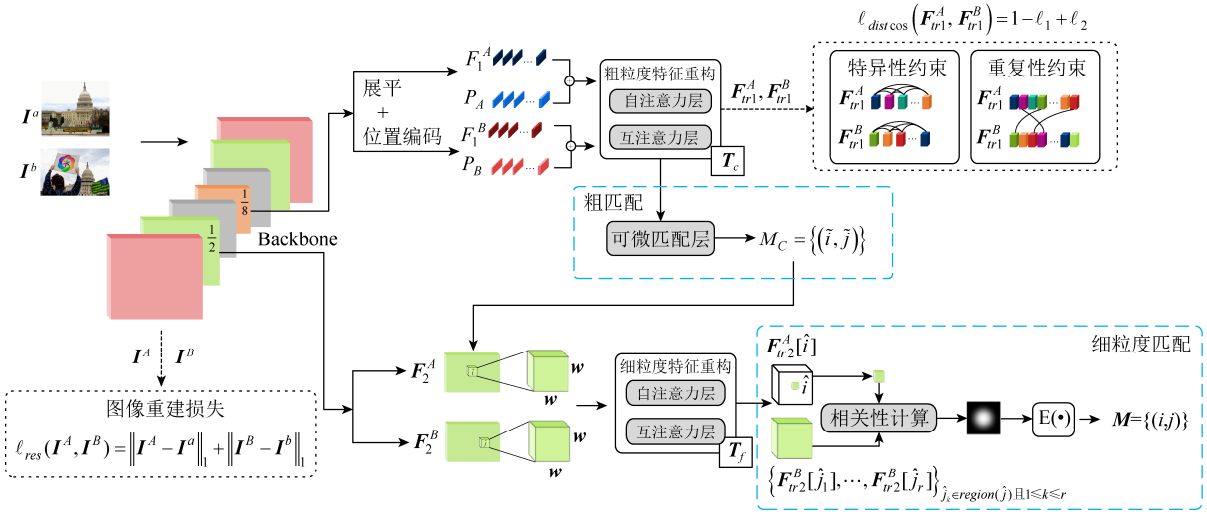


图 1 总体框架

Fig. 1 Overview of framework

本文的贡献在于创新性地提出了一种无检测器图像特征匹配网络，能够有效地处理不同质量的图像。对使用自注意力和互注意力的 Transformer[3] 重构得到的特征进行约束，提高特征描述子的重复性和特异性，从而实现更好的匹配效果。同时，在网络的 Backbone 解码阶段增加图像重建层，有助于更好地适应不同质量图像的特征匹配。

1　相关工作

1.1　基于检测器的图像特征匹配

基于检测器的方法是图像特征匹配的主要方法。在深度学习之前，图像特征匹配方法利用手工定义的局部特征取得了良好的表现，例如 SIFT[4] 和 ORB[5]。ORB 改进了 FAST 检测子[6]不具有方向性的问题，并采用速度极快的二进制描述子 BRIEF[7]加速图像特征提取环节。ORB 特征由关键点和描述子 2 部分组成，ORB 的关键点称为“Oriented FAST”，是一种改进的 FAST 角点。ORB 特征用于视觉 SLAM 系统 ORB-SLAM[8]有效提高了系统的定位和建图性能，但 ORB 在弱纹理、图像内容重复、视点变化较大等情形下不能够提取出足够多且准确的关键点。使用基于深度学习的方法可以显著提高视点及照明变化较大等情况下的特征表示能力，LIFT[9]和 MagicPoint[10]最早成功地实现了基于深度学习的局部特征表示和基于深度学习的 SLAM。SuperPoint[11]在 MagicPoint 的基础之上，引入 Homographic Adaptation[11]机制，创建了一个大规模的 pseudo-ground truth 数据集，利用关键点检测器而非人工标注做为监督。其缺点是，让当前模型模仿另一个关键点检测器导致其无法发现潜在的最佳关键点。虽然 SuperPoint 相比传统方法显著提升了性能，但无法在图像弱纹理区域提取足够数量的关键点。

上述方法均使用最近邻搜索算法计算关键点间的匹配，而 SuperGlue[12]提出一种基于学习的局部特征匹配方法。输入 2 幅图像中的关键点以及描述子(手工特征或深度学习特征均可)，SuperGlue 在图神经网络(graph neural networks，GNN)[13]基础上引入一种基于注意力的上下文聚合机制学习关键点的匹配。由于通过数据驱动方法学习特征匹配的先验，SuperGlue 实现了令人印象深刻的性能。SuperGlue 开创了特征匹配的新技术，使用自注意力和互注意力学习上下文以及基于最优传输实现匹配的思想被后续相关研究灵活应用。然而，作为一种依赖检测器的方法，SuperGlue 关注的范围仅限于由检测器检测到的关键点。

1.2　无检测器的图像特征匹配

无检测器的图像特征匹配方法不使用特征检测器，直接生成稠密特征描述子或稠密特征匹配。NCNet[14]是一种以端到端的方式直接学习稠密匹配关系的方法，构造 4D cost volumes[15]并枚举图像之间所有可能的匹配，基于网络识别可靠的匹配对并过滤不可靠的匹配，但是 4D 卷积的感受野仍然受限于每个匹配对的邻域。受 SuperGlue 的启发，LoFTR 使用具有自注意和互注意的 Transformer 处理基于卷积神经网络(convolutional neural networks，CNN)提取的稠密局部特征，Transformer 的全局感受野和位置编码使得学习到的特征与上下文和位置相关。LoFTR 在弱纹理区域能够得到比较好的匹配，但仍存在一定的错误匹配。此外，对于模糊图像匹配，LoFTR 表现不佳。本文充分考虑参与匹配的特征点的重复性和特异性，以提高图像特征匹配的精度和鲁棒性。

目前少有模糊图像特征匹配的研究，且没有能够适应不同质量图像的特征匹配方法。本文提出的图像特征匹配方法能够适应不同质量图像，可以应对机器人或无人车高速行驶时传感器故障或传感器发生快速位移可能导致的图像运动模糊情况。

2　方法

本文总体框架如图 1 所示。首先，提取输入图像的特征，并进行图像重建，以适应不同质量图像的特征匹配。其次，基于粗粒度特征重构模块对初步提取的特征进行特征重构，并约束重构的特征具有重复性和特异性。最后，基于粗粒度特征和细粒度特征依次执行由粗到细的匹配，最终得到匹配点对集。

2.1　特征提取网络

图 1 所示的 Backbone 为特征提取网络，是基于 ResNet-18[16]和三层 FPN[17]构建的。与原始 ResNet-18 不同的是，第一个卷积层的通道数为 128，后续 3 个 Block 的通道数分别为 128，196 和 256。FPN 具有 $\boldsymbol{P}_1$ 到 $\boldsymbol{P}_3$ 3 个层级，$\boldsymbol{P}_3$ 的特征 $\boldsymbol{F}_1^A$ 和 $\boldsymbol{F}_1^B$ 是粗粒度特征重构模块的输入，$\boldsymbol{F}_1^A$ 和 $\boldsymbol{F}_1^B$ 表示大小为 1/8 原始图像尺寸的粗级特征；$\boldsymbol{P}_1$ 的特征 $\boldsymbol{F}_2^A$ 和 $\boldsymbol{F}_2^B$ 是细粒度特征重构模块的输入，$\boldsymbol{F}_2^A$ 和 $\boldsymbol{F}_2^B$ 表示大小为 1/2 原始图像尺寸的精细级特征。

特征提取网络的最后一层为反卷积层，用于重建图像，使用图像重建损失约束网络学习更鲁棒的特征表示，以适应不同质量图像的特征匹配。图像重建损失函数为

$$\ell_{res}(\boldsymbol{I}^A,\boldsymbol{I}^B)=\|\boldsymbol{I}^A-\boldsymbol{I}^a\|_1+\|\boldsymbol{I}^B-\boldsymbol{I}^b\|_1 \quad (1)$$

其中，$\boldsymbol{I}^A$ 和 $\boldsymbol{I}^B$ 为重建之后的图像；$\boldsymbol{I}^a$ 和 $\boldsymbol{I}^b$ 为源图像。

2.2 粗粒度特征重构及匹配

将 Backbone 输出的、尺寸为 $h\times w\times c$ 的粗级特征展平为 $c\times hw$ 的特征图 $\boldsymbol{F}_1^A$ 和 $\boldsymbol{F}_1^B$ 并基于像素坐标进行位置编码[2]，编码后的局部特征向量输入粗粒度特征重构模块进行特征重构，重构之后的特征输入可微匹配层得到粗匹配点对集。

2.2.1 粗粒度特征重构

粗粒度特征重构模块使用 Transformer 编码器，采取自注意和互注意的方式对特征进行重构，其中自注意力层和互注意力层交错 T_c 次。自注意力层的输入来自同一个特征图 $\boldsymbol{F}_1^A$ 或 $\boldsymbol{F}_1^B$，互注意力层的输入分别来自 2 幅特征图 $\boldsymbol{F}_1^A$ 和 $\boldsymbol{F}_1^B$。$\boldsymbol{F}_{tr1}^A$ 和 $\boldsymbol{F}_{tr1}^B$ 表示大小为原始图像尺寸 1/8 的特征图 $\boldsymbol{F}_1^A$ 和 $\boldsymbol{F}_1^B$ 输入到粗粒度特征重构模块中得到的重构特征。

粗粒度特征重构模块使用的 Transformer 编码器结构如图 2 所示，Multi-head attention 的输入向量通常被命名为查询、键和值。输入向量首先变换为 3 个不同的向量，即：查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。注意力层可表示为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Soft\,max}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V} \quad (2)$$

注意力层通过测量查询向量和关键向量之间的相似性来选择相关信息，其输出向量是由相似度评分加权的值向量之和。因此，如果相似度较高，可从值向量中提取较多信息。

对重构之后的特征进行重复性和特异性约束，即

$$\ell_{dist\cos}\left(\boldsymbol{F}_{tr1}^A,\boldsymbol{F}_{tr1}^B\right)=1-\ell_1+\ell_2 \quad (3)$$

$$\ell_1=\sum_{(\tilde{i},\tilde{j})\in\boldsymbol{M}_c^{gt}} dist\cos\left(\boldsymbol{F}_{tr1}^A[\tilde{i}],\boldsymbol{F}_{tr1}^B[\tilde{j}]\right)$$

$$\ell_2=\sum_{\tilde{i},\tilde{i}'\in I_A} dist\cos\left(\boldsymbol{F}_{tr1}^A[\tilde{i}],\boldsymbol{F}_{tr1}^A[\tilde{i}']\right)+\sum_{\tilde{j},\tilde{j}'\in I_B} dist\cos\left(\boldsymbol{F}_{tr1}^B[\tilde{j}],\boldsymbol{F}_{tr1}^B[\tilde{j}']\right)$$

其中，$\boldsymbol{F}_{tr1}^*[\bullet]$ 为特征重构模块输出的某像素的特征向量；$dist\cos$ 为特征之间的余弦距离；$\boldsymbol{M}_c^{gt}$ 为真实匹配的集合，是由真实的相机位姿以及与输入图像对应的深度图计算得到的；ℓ_1 为重复性约束，约束图像 A 和 B 中匹配点的特征的余弦距离尽可能近；ℓ_2 为特异性约束，约束图像 A 或 B 中某像素特征和其他像素特征的余弦距离尽可能远。

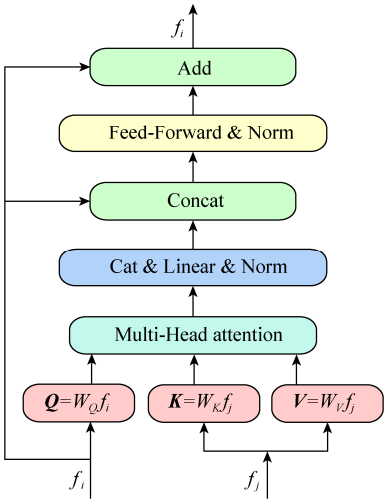


图 2 Transformer 编码器层

Fig. 2 Transformer encoder layer

式(3)定义的损失函数约束网络学习更高质量的像素特征，要求 2 幅图像中具有真实匹配关系的点的特征更加相似，同时保证同一幅图像中不同像素的特征之间更加具有差异性。高质量的像素特征可以避免由于相邻像素特征高度相似而导致误匹配，且有利于更准确地找到在另一幅图像中的匹配点。

2.2.2 粗匹配

如图 1 所示，粗粒度特征重构模块的输出 $\boldsymbol{F}_{tr1}^A$ 和 $\boldsymbol{F}_{tr1}^B$ 输入可微匹配层，计算得到得分矩阵 $\boldsymbol{S}$，$\boldsymbol{S}[i,j]=Corr\left(F_{tr1}^A[i],F_{tr1}^B[j]\right)$。与 LoFTR 一样，本文采用双 Softmax[18-19]算子，在 $\boldsymbol{S}$ 的 2 个维度上应用 Softmax，得到匹配概率为

$$\boldsymbol{P}_c[i,j]=\mathrm{Softmax}(\boldsymbol{S}[i,\cdot])_j\times\mathrm{Softmax}(\boldsymbol{S}[\cdot,j])_i \quad (4)$$

基于 $\boldsymbol{P}_c$ 选择高于置信阈值 θ_c 的匹配，然后使用相互最近邻算法得到粗粒度匹配点对集 $\boldsymbol{M}_c=\{(\tilde{i},\tilde{j})\}$。粗粒度匹配的损失函数为交叉熵损失，即

$$\ell_c=-\frac{1}{|\boldsymbol{M}_c^{gt}|}\sum_{(\tilde{i},\tilde{j})\in M_c^{gt}} FL(\boldsymbol{P}_c(\tilde{i},\tilde{j}))\log\boldsymbol{P}_c(\tilde{i},\tilde{j}) \quad (5)$$

其中，FL 为 focal loss，用于解决匹配对和非匹配对之间的数量不均衡问题。

2.3　细粒度特征重构及匹配

对于每一个粗粒度匹配点对 $(\tilde{i},\tilde{j})$，首先在细粒度特征图 $\boldsymbol{F}_2^A$ 和 $\boldsymbol{F}_2^B$ 上确定其位置 $(\hat{i},\hat{j})$，然后裁剪 2 组分别以 $\hat{i}$ 和 $\hat{j}$ 为中心，大小为 $w\times w$ 的局部窗口，使用细粒度特征重构模块中的自注意力层和互注意力层将窗口内的特征变换 T_f 次，生成以 $\hat{i}$ 和 $\hat{j}$ 为中心的局部特征图 $\boldsymbol{F}_{tr2}^A$ 和 $\boldsymbol{F}_{tr2}^B$。

基于特征图 $\boldsymbol{F}_{tr2}^A$ 和 $\boldsymbol{F}_{tr2}^B$，计算 $\hat{i}$ 的特征向量 $\boldsymbol{F}_{tr2}^A[\hat{i}]$ 和以 $\hat{j}$ 为中心的局部窗口内所有像素对应的特征向量 $\left\{\boldsymbol{F}_{tr2}^B[\hat{j}_1],\cdots,\boldsymbol{F}_{tr2}^B[\hat{j}_r]\right\}_{\hat{j}_k\in region(\hat{j})\text{且}1\leqslant k\leqslant r}$ 的相关性。基于相关性生成热图，该热图表示 $\hat{i}$ 与 $\hat{j}$ 邻域中每个像素的匹配概率，通过计算概率分布期望 E(·)得到最终的精细匹配 $\boldsymbol{M}=\{(i,j)\}$。

3　实验

本文进行了图像特征匹配的单应性估计和相对位姿估计实验，并验证了不同质量图像的特征匹配的有效性。

3.1　实验设置

本文在 ScanNet 数据集上训练室内模型，在 MegaDepth 上训练室外模型。设置粗粒度特征重构模块中 T_c 为 4，置信度分数阈值 θ_c 为 0.2，细粒度特征重构模块中 T_f 为 1，窗口宽度 w 为 5。对于室内数据集 ScanNet，使用初始学习率设置为 0.004 的 Adam 模型在 2 块 GTX 3090 上训练，共训练 30 个 epoch，batch 大小为 8。对于室外数据集 MegaDepth，初始学习率设置为 0.003，共训练 30 个 epoch，batch 大小为 4。

3.2　单应性估计

3.2.1　数据集及评价指标

在单应性估计实验中，与 LoFTR[2]相同，本文使用 HPatches[20]数据集，其包含 52 个显著照明变化下的序列和 56 个视点变化较大的序列。

在每个测试序列中，一个参考图像与其余 5 个图像配对，所有图像的短边的尺寸调整到 480。对所有图像对，本文使用在 MegaDepth[21]上训练的模型提取匹配点集，使用 OpenCV 中的 RANSAC 计算单应矩阵 $\boldsymbol{H}$[22]。

为了与产生不同数量匹配的方法进行公平比较，本文计算估计的单应矩阵扭曲图像和真实的单应矩阵扭曲图像之间的角误差，报告阈值分别为 3，5 和 10 像素下的角误差累积曲线下的面积。

3.2.2　与前沿方法的比较与分析

基于检测器的图像特征匹配方法，包括 R2D2[22]，D2Net[23]，DISK[24]和 SuperGlue[12]；无检测器的图像特征匹配方法，包括 DRC-Net[25]和 LoFTR[2]，本文将匹配点对数量设置为 1 K。表 1 中本文方法在不同角误差阈值下的单应性估计 AUC 明显优于其他方法。LoFTR 最早引入基于 Transformer 的自注意力与互注意力用于特征重构，其单应性估计 AUC 明显优于其他已有方法。本文在 LoFTR 基础之上对网络进行修改，对重构之后特征进一步约束，使得单应性估计 AUC 达到最佳。

表 1　HPatches 上单应性估计

Table 1　Homography estimation on HPatches

类别	方法	单应性估计 AUC		
		@3px	@5px	@10px
有检测器	D2Net+NN	23.2	35.9	53.6
	R2D2+NN	50.6	63.9	76.8
	DISK+NN	52.3	64.9	78.9
	SP+ SuperGlue	53.9	68.3	81.7
无检测器	DRC-Net	50.6	56.2	68.3
	LoFTR	65.9	75.6	84.6
	Ours	**66.8**	**76.9**	**86.1**

注：加粗数据为最优值

3.3　相对位姿估计

3.3.1　数据集及评价指标

使用 ScanNet 和 MegaDepth 数据集证明本文方法用于位姿估计的有效性。

ScanNet 为室内场景数据集，包含 1613 个带有真实位姿和深度图的单目序列。与 LoFTR[2]一样，本文选用 230 M 图像对进行训练，选择其中的 1 500 个测试对进行评估，所有的图像和深度图尺寸均调整为 640×480。ScanNet 数据集包含大量无纹理区域图像对。

MegaDepth 由 196 个不同户外场景的图像组成，数据集提供了来自 COLMAP[26]的稀疏重建和由双目多视图计算得到的深度图。在和前沿方法比较时，与 LoFTR[2]一样随机抽取 1 500 对图像进行公平比较。训练和测试阶段需要调整图像大小，训练阶段图像的长边调整为 840，验证阶段图像的长边调整为 1 200。MegaDepth 数据集的特点是图像的视点变化较大且场景内容重复较多。相对位姿估计误差定义为旋转和平移的角度最大误差。

3.3.2 与前沿方法的比较与分析

表 2 和表 3 展示了位姿估计误差 AUC 的值，本文方法取得了最优位姿估计精度。SP[11]+Superglue 和 DRC-Net 仅考虑特征点周围局部图像块的信息，而 LoFTR 基于 Transformer 提取更大尺度区域的上下文信息，因此相对位姿估计的误差较小。本文方法计算得到的像素点特征包含更丰富的全局信息，同时通过对像素点特征实施重复性和特异性约束促使网络学习更高质量的像素特征，

表 2 室内数据集 ScanNet 上的相对位姿估计

Table 2 Relative pose estimation on indoor dataset ScanNet

类别	方法	位姿估计 AUC		
		@5°	@10°	@20°
有检测器	SP[11]+Superglue	16.16	33.81	51.84
无检测器	DRC-Net	7.69	17.93	30.49
	LoFTR	22.06	40.80	57.96
	Ours	**22.87**	**41.75**	**59.10**

注：加粗数据为最优值

表 3 室外数据集 MegaDepth 上的相对位姿估计

Table 3 Relative pose estimation on outdoor dataset MegaDepth

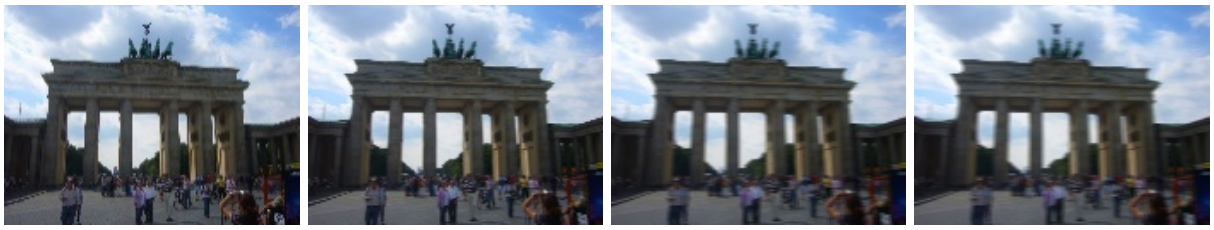
类别	方法	位姿估计 AUC		
		@5°	@10°	@20°
有检测器	SP+Superglue	42.18	61.16	75.96
无检测器	DRC-Net	27.01	42.96	58.31
	LoFTR	52.81	69.19	81.18
	Ours	**53.63**	**70.20**	**83.56**

注：加粗数据为最优值

使得在弱纹理区域能够得到可靠的匹配点对，从而使相对位姿估计的误差更小。

3.4 模糊图像特征匹配

在 SLAM 和位姿估计任务中对模糊图像的特征匹配研究较少且缺少相应的数据集，本文构建了包含不同模糊程度图像的新数据集 MegaDepth-B，并基于构建的新数据集验证提出方法在模糊图像匹配任务上的有效性。本文构建模糊数据的方法是利用不同尺寸的模糊核对图像进行卷积，图 3 所示为生成的模糊图像。



(a) (b) (c) (d)

图 3 模糊图像示例((a)原图像；(b)模糊核 5×5；(c)模糊核 12×12；(d)模糊核 24×24)

Fig. 3 Examples of blurred image ((a) Original image; (b) Blurring kernel 5×5; (c) Blurring kernel 12×12; (d) Blurring kernel 24×24)

本文对数据集 MegaDepth 中的图像进行模糊处理，即采用大小为 5×5，12×12 和 24×24 的模糊核对图像进行卷积得到 3 种不同模糊程度的图像。表 4 给出了不同模糊程度下 LoFTR[2]和本文方法的位姿估计实验结果，第 2，4 和 6 行表明本文方法基于不同模糊程度图像的位姿估计 AUC 均较 LoFTR 有很大程度提升，且基于不同模糊程度图像的位姿估计 AUC 变化ΔAUC 小于 LoFTR，进一步证明了本文方法对模糊图像特征匹配的有效性。

本文方法对模糊图像的特征匹配显著有效，原因在于采用了图像重建损失约束。该约束利用清晰图像作为真值监督模糊图像的重建，从而促使网络学习到更加鲁棒的特征表示。

表 4 基于不同模糊程度图像的位姿估计对比

Table 4 Comparison of pose estimation using images with different blurriness

方法	模糊核	位姿估计 AUC		
		@5°	@10°	@20°
LoFTR	5×5	40.63	56.70	70.53
Ours	5×5	**44.60**	**63.50**	**76.52**
LoFTR	12×12	32.37	47.32	61.68
Ours	12×12	**41.10**	**59.5**	**70.63**
LoFTR	24×24	18.86	31.86	47.18
Ours	24×24	**32.68**	**45.20**	**57.24**

注：加粗数据为最优值

3.5 消融实验

为了验证本文提出的图像重建模块和粗粒度特征重构模块中特异性和重复性约束的有效性，

本文在 MegaDepth 数据集和 MegaDepth-B 数据集上进行了消融实验，实验结果见表 5。具体实验如下：

(1) 验证图像重建模块的有效性时，在网络框架中去除粗粒度特征重构模块的重复性和特异性约束。实验结果表明，本文提出的图像重建模块可有效提高基于不同质量图像的相机位姿估计的精度，尤其是基于模糊图像的位姿估计精度有明显提升。

(2) 验证粗粒度特征重构模块的重复性和特异性约束的有效性时，在网络框架中去除了图像重建模块。实验结果表明，粗粒度特征重构模块的重复性和特异性约束对清晰图像和模糊图像的特征匹配同样有效。此外，表 5 的实验结果表明，无论是清晰图像还是模糊图像，图像重建模块和重复性及特异性约束共存时，相机位姿估计精度达到最优。

表 5　消融实验

Table 5　Ablation Experiment

数据集	图像重建	重复性和特异性约束	位姿估计 AUC		
			@5°	@10°	@20°
MegaDepth	√	√	**53.63**	**70.20**	**83.56**
	√	-	52.88	69.30	81.18
	-	√	53.58	70.16	83.47
	-	-	52.81	69.19	81.18
MegaDepth-B (模糊核 12×12)	√	√	**41.10**	**59.5**	**70.63**
	√	-	40.56	58.4	69.03
	-	√	33.96	49.12	63.98
	-	-	32.37	47.32	61.68

注：加粗数据为最优值

3.6　可视化结果

图 4 展示了在室外数据集 MegaDepth 上的可视化结果，第一行、第二行分别为 LoFTR[2]和本文方法的可视化结果。红色的线表示极线误差[2]超过 1×10^{-4} 的匹配点对，很显然本文方法的正确匹配点对更多，且相对位姿估计的角度误差远远小于 LoFTR。如图 4(c)所示，当图像质量较差、模糊程度较高(模糊核 24×24)时 LoFTR 表现更差，几乎不能得到正确的匹配，其中旋转的角度误差达到了 51.55°，而本文方法只有 26.54°。在极线误差阈值为 1×10^{-4} 时，图 4 所示本文方法得到的正确匹配点对更多(红色的线相对较少，绿色的线相对较多)，原因是图像重建模块的图像重建损失约束网络学习更鲁棒的特征表示，粗粒度特征重构模块的重复性和特异性损失约束网络学习更高质量的像素特征。

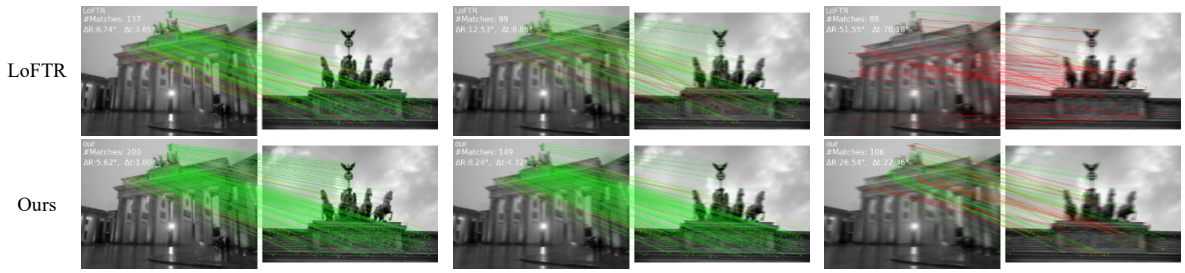


图 4　MegaDepth 数据集上可视化结果((a) MegaDepth 中的清晰图像；(b)模糊程度适中(模糊核 12×12)；(c)图像质量较差、模糊程度较高(模糊核 24×24))

Fig. 4　Visualization results on MegaDepth dataset ((a) Clear images in MegaDepth; (b) Moderate blurring (blurring kernel 12×12); (c) Poor image quality and high blurring (blurring kernel 24×24))

4　结束语

本文在现有图像特征匹配框架 LoFTR 的基础上，引入重复性约束和特异性约束，增强了同一幅图像内像素特征的区分度，并使不同图像的可匹配点的特征具有更强的相似性。同时，在网络的解码阶段增加图像重建层，提高了网络学习到的特征表示的鲁棒性。在室内数据集 ScanNet 和室外数据集 MegaDepth 上的单应性估计和相对位姿估计实验结果表明，本文提出的重复性约束和特异性约束对于图像特征匹配具有显著的效果。基于不同质量图像数据的位姿估计实验结果验证了本文方法的鲁棒性。此外，在不同质量图像上的消融实验表明，本文提出的重复性和特异性约束以及图像重建模块对于图像特征匹配具有较好的效果。

本文的网络模型规模较大，不利于部署到资源有限的应用场景，下一步将重点考虑优化网络结构。

参考文献 (References)

[1] 吴凡, 宗艳桃, 汤霞清. 视觉 SLAM 的研究现状与展望[J]. 计算机应用研究, 2020, 37(8): 2248-2254.
WU F, ZONG Y T, TANG X Q. Research status and prospect of vision SLAM[J]. Application Research of Computers, 2020, 37(8): 2248-2254 (in Chinese).

[2] SUN J M, SHEN Z H, WANG Y A, et al. LoFTR: detector-free local feature matching with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 8922-8931.

[3] KATHAROPOULOS A, VYAS A, PAPPAS N, et al. Transformers are RNNs: fast autoregressive transformers with linear attention[EB/OL]. [2022-05-11]. https://arxiv.org/abs/2006.16236.

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

[5] RUBLEE E, RABAUD V, KONOLIGE K, et al. ORB: an efficient alternative to SIFT or SURF[C]//2011 International Conference on Computer Vision. New York: IEEE Press, 2011: 2564-2571.

[6] ROSTEN E, DRUMMOND T. Machine learning for high-speed corner detection[C]//The 9th European Conference on Computer Vision - Volume Part I. New York: ACM, 2006: 430-443.

[7] CALONDER M, LEPETIT V, STRECHA C, et al. Brief: binary robust independent elementary features[C]//European Conference on Computer Vision. Heidelberg: Springer, 2010: 778-792.

[8] MUR-ARTAL R, TARDÓS J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262.

[9] YI K M, TRULLS E, LEPETIT V, et al. LIFT: Learned Invariant Feature Transform[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 467-483.

[10] ETONE D, MALISIEWICZ T, RABINOVICH A. Toward geometric deep SLAM[EB/OL]. [2022-05-16]. https://arxiv.org/abs/1707.07410.

[11] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE Press, 2018: 337:1-337:12.

[12] SARLIN P E, DETONE D, MALISIEWICZ T, et al. Superglue: learning feature matching with graph neural networks[C]//2020 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 4938-4947.

[13] 白铂, 刘玉婷, 马驰骋, 等. 图神经网络[J]. 中国科学: 数学, 2020, 3: 367-384.
BAI B, LIU Y T, MA C C, et al. Graph neural network[J]. Science in China: Mathematics, 2020, 3: 367-384 (in Chinese).

[14] ROCCO I, ARANDJELOVIĆ R, SIVIC J. Efficient Neighbourhood Consensus Networks via Submanifold Sparse Convolutions[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 605-621.

[15] YANG G, RAMANAN D. Volumetric correspondence networks for optical flow[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2019: 794-805.

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 2117-2125.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 770-778.

[18] ROCCO I, CIMPOI M, ARANDJELOVIĆ R, et al. Neighbourhood consensus networks[EB/OL]. [2022-05-16]. https://arxiv.org/abs/1810.10510.

[19] TYSZKIEWICZ M J, FUA P, TRULLS E. DISK: learning local features with policy gradient[C]//The 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 14254-14265.

[20] BALNTAS V, LENC K, VEDALDI A, et al. HPatches: a benchmark and evaluation of handcrafted and learned local descriptors[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 5173-5182.

[21] LI Z Q, SNAVELY N. MegaDepth: learning single-view depth prediction from Internet photos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 2041-2050.

[22] REVAUD J, WEINZAEPFEL P, DE SOUZA C, et al. R2D2: repeatable and reliable detector and descriptor[EB/OL]. [2022-05-16]. https://arxiv.org/abs/1906.06195.

[23] DUSMANU M, ROCCO I, PAJDLA T, et al. D2-net: a trainable cnn for joint detection and description of local features[EB/OL]. [2022-05-16]. https://arxiv.org/abs 1905. 03561.

[24] TYSZKIEWICZ M J, FUA P, TRULLS E. DISK: learning local features with policy gradient[C]//The 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 14254-14265.

[25] LI X H, HAN K, LI S D, et al. Dual-resolution correspondence networks[C]//The 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 17346-17357.

[26] SCHÖNBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 4104-4113.