文章编号：2095-2163(2022)09-0198-06 中图分类号：TP391.4 文献标志码：A

# 基于改进的 ResNet 与 IMU 位姿图像特征描述子

陈守刚，张伟伟，赵 波
（上海工程技术大学 机械与汽车工程学院，上海 201620）

**摘 要**：基于深度学习的图像特征描述子，是许多 3D 视觉任务的重要组成部分，但现有的基于深度学习的图像特征描述子框架，通常需要特征点之间的真实对应关系来进行训练，而要想大规模获取这些对应的特征点却具有很大的挑战性。本文提出了一种新的弱监督学习框架，该框架只需从与图像相关联的惯性测量单元位姿中学习特征描述子。基于此，本文构造了新的损失函数，该函数利用 IMU 位姿所给定的对极约束，方法稳定且高效。因为本方法不需要特征点之间的真实对应关系，所以在庞大且多样化的数据集上训练效果更好，为更具有区分性的局部特征描述子提供了可能。本文将学习到的特征描述子称为 POSE 描述子，经过严格的监督训练，POSE 描述子比之前基于完全监督的特征描述子更好，且数量和匹配度均有所提高。

**关键词**：深度学习；IMU 位姿；特征描述子；对极约束

## Images feature descriptors based on improved ResNet and IMU

CHEN Shougang, ZHANG Weiwei, ZHAO Bo
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

**【Abstract】** Images feature descriptors based on deep learning are an important part of many 3D vision tasks, and they have developed rapidly in recent years. However, the existing deep learning-based images feature descriptors usually require real correspondence between feature points for training, and it is very challenging to obtain these corresponding feature points. This paper proposes a new weakly supervised learning framework, which only needs to learn feature descriptors from the IMU pose associated with the picture. Based on this, this paper constructs a new loss function, which utilizes the epipolar constraint given by the IMU pose, and the method is stable and efficient. Because this method does not require the true correspondence between feature points, this framework provides the possibility to train better and more discriminative local feature descriptors on larger and more diversified data sets. In this paper, the learned feature descriptor is called POSE descriptors. After strict supervision training, the POSE descriptor is better than the previous feature descriptors based on full supervision, and the number of matched correspondences and the matching degree are improved.

**【Key words】** deep learning; IMU pose; feature descriptor; antipolar constraint

## 0 引 言

寻找图像局部特征稳定且高效的对应关系，是计算机视觉任务的基本组成部分。例如，基于多视图几何三维重建[1]（Structure from motion，SFM）和同步定位与建图[2]（Simultaneous Localization and Mapping，SLAM）都需要稳定且区分度高的特征描述子。目前，基于深度学习的图像特征描述子[3-4]（如：SuperPoint、LF-Net），展现出比人工设计的特征描述子具有更好的性能。然而，一些研究表明，当把 SuperPoint 等特征描述子应用于有遮挡的现实世界时，存在泛化能力弱的问题。存在这种局限性的一个重要原因，就是无法获取图像对之间特征点真实的对应关系。之前，许多方法都采用 SFM 数据集作为替代方案，但这些数据集提供的匹配特征点并不是真实的对应关系。

针对上述问题，本文方法不要求特征点之间具有严格的对应关系，仅从图像对之间的相对相机位姿中学习特征描述子，就可以通过各种基于非视觉的传感器，例如惯性测量单元（Inertial Measurement Unit，IMU）获得相机姿态；通过减少特征点之间严格匹配的要求，可以在多样化的数据集上学习更好的特征描述子，解决了特征点学习获取训练数据的困难。

然而，由于不能基于相机姿势来构造损失函数，现有的深度学习方法不能直接利用相机位姿作为约束。本文的主要贡献：提出了一种新的框架，将图像对之间的 IMU 位姿转换为对匹配点之间的像素位

作者简介：陈守刚（1995-），男，硕士研究生，主要研究方向：计算机视觉、视觉 SLAM；张伟伟（1987-），男，博士，副教授，主要研究方法：自动驾驶、计算机视觉、深度学习；赵 波（1962-），女，硕士，副教授，主要研究方向：智能汽车技术、汽车传动技术、汽车轻量化设计与制造。

收稿日期：2021-11-13

置的对极约束作为监督约束条件(如图 1 所示)；匹配点的位置相对于用于训练的特征描述子是可区分的；为了进一步降低计算成本并加快训练速度，使用了从粗到精的匹配方案，以较低的分辨率计算对应关系，以更精细的比例进行局部优化。

(a) 位姿一　　(b) 位姿二

**图 1　位姿一和位姿二图像匹配对之间的查询点和预测点之间的对应关系**

**Fig. 1　Correspondence between query points and prediction points between pose-1 and pose-2 image matching pairs**

## 1　IMU 器件测量与运动学模型

### 1.1　陀螺仪与加速度计的测量模型

陀螺仪的测量模型为：

$$\tilde{\omega}_{wb}^{b}(t)=\omega_{wb}^{b}(t)+b_g(t)+\eta_g(t) \tag{1}$$

其中，$b_g$ 是随时间缓慢变化的偏差，$\eta_g$ 是白噪声。模型利用了静态世界假设[5]，即重力加速度不发生变化。

加速度计的测量模型为：

$$f^{b}(t)=R_b^{wT}(a^{w}-g^{w})+b_a(t)+\eta_a(t) \tag{2}$$

其中，$b_a$ 是随着时间缓慢变化的偏差，$\eta_a$ 是白噪声。

### 1.2　陀螺仪与加速度计的运动模型

运动模型的微分方程形式为：

$$\begin{cases}\dot{\hat{R}}_b^w=R_b^w(\omega_{wb}^b)^{\wedge}\\ \dot{v}^w=a^w\\ \dot{p}^w=v^w\end{cases} \tag{3}$$

使用欧拉积分[6]、即三角积分可以得到运动方程的离散形式：

$$\begin{cases}R_{b(t+\Delta t)}^w=R_{b(t)}^w Exp(\omega_{wb}^b(t)\cdot\Delta t)\\ v^w(t+\Delta t)=v^w(t)+a^w(t)\cdot\Delta t\\ p^w(t+\Delta t)=p^w(t)+v^w(t)\cdot\Delta t+\frac{1}{2}a^w(t)\cdot\Delta t^2\end{cases} \tag{4}$$

其中，$\omega_{wb}^b(t)$ 表示 $t$ 时刻“角速度矢量”在 $b$ 坐标系下的坐标，$\omega_{wb}^b(t)\cdot\Delta t$ 表示“旋转矢量”在 $b$ 坐标系下的坐标。

### 1.3　测量与运动混合模型

为了使符号简明，对符号进行重新定义为：

$R(t)\doteq R_{b(t)}^w$；$\omega(t)\doteq\omega_{wb}^b(t)$；$f(t)\doteq f^b(t)$；$v(t)\doteq v^w(t)$；$p(t)\doteq p^w(t)$；$g\doteq g^w$

将测量模型代入运动方程：

$$\begin{cases}R(t+\Delta t)=R(t)\cdot Exp(\omega(t)\cdot\Delta t)=R(t)\cdot Exp((\hat{\omega}(t)-b_g(t)-\eta_{gd}(t))\cdot\Delta t)\\ v(t+\Delta t)=v(t)+a^w(t)\cdot\Delta t=v(t)+R(t)\cdot(\hat{f}(t)-b_a(t)-\eta_{ad}(t))\cdot\Delta t+g\cdot\Delta t\\ p(t+\Delta t)=p(t)+v(t)\cdot\Delta t+\frac{1}{2}a^w(t)\cdot\Delta t^2=p(t)+v(t)\cdot\Delta t+\frac{1}{2}[R(t)\cdot(\hat{f}(t)-b_a(t)-\eta_{ad}(t))+g]\cdot\Delta t^2=p(t)+v(t)\cdot\Delta t+\frac{1}{2}g\cdot\Delta t^2+\frac{1}{2}R(t)\cdot(\hat{f}(t)-b_a(t)-\eta_{ad}(t))\cdot\Delta t^2\end{cases} \tag{5}$$

其中，噪声项采用 $\eta_{gd}$ 和 $\eta_{ad}$($d$ 表示 discrete)，故与连续噪声项 $\eta_g$ 和 $\eta_a$ 是不同的。离散噪声与连续噪声的协方差有如下关系：

$$\begin{cases}Cov(\eta_{gd}(t))=\frac{1}{\Delta t}Cov(\eta_g(t))\\ Cov(\eta_{ad}(t))=\frac{1}{\Delta t}Cov(\eta_a(t))\end{cases} \tag{6}$$

进一步假设 $\Delta t$ 恒定、即采样频率不变，每个离散时刻由 $k=0,1,2,\cdots$ 表示，上述的 3 个离散运动方程可进一步简化为：

$$\begin{cases}R_{k+1}=R_k\cdot Exp((\hat{\omega}_k-b_k^g-\eta_k^{gd})\cdot\Delta t)\\ v_{k+1}=v_k+R_k\cdot(\hat{f}_k-b_k^a-\eta_k^{ad})\cdot\Delta t+g\cdot\Delta t\\ p_{k+1}=p_k+v_k\cdot\Delta t+\frac{1}{2}g\cdot\Delta t^2+\frac{1}{2}R_k\cdot(\hat{f}_k-b_k^a-\eta_k^{ad})\cdot\Delta t^2\end{cases} \tag{7}$$

## 2 ResNet 简介

### 2.1 ResNet 架构

随着网络的加深,神经网络梯度消失,出现了训练集准确率下降的现象。为了解决上述问题,He 等人[7]在 2016 年提出 ResNet 深度学习模型,该模型的最大特性就是网络层深度可以无限叠加,而不会出现梯度消失的问题。该网络的提出,说明网络的深度对许多计算机视觉识别任务至关重要,在 ImageNet 数据集上的错误率仅为 3.57%,获得 ILSVRC 2015 中图像分类任务的第一名。ResNet 在 COCO 目标检测数据集上获得了 28%的相对改进。

ResNet 的残差模型如图 2 所示。图 2 中,$x$ 表示输入的 feature mapping,$F(x)$ 为残差,$F(x)+x$ 表示下一层网络的输入,该网络特征可以有效确保梯度不会消失。

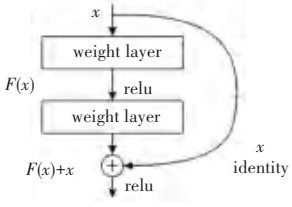


**图 2 ResNet 的残差模型**

**Fig. 2 Residual model of ResNet**

虽然 ResNet 网络的层数可以无限叠加,但是更深的网络会导致训练的复杂度变大。综合现有研究结果可知,ResNet 采用 18、34、50、101、152 这 5 种深度的网络结构较为普遍。本文所提出的方法就是基于 ResNet50 网络结构,如图 3 所示。

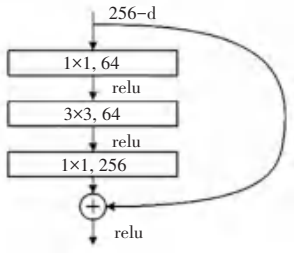


**图 3 ResNet50 的残差模型**

**Fig. 3 Residual model of ResNet50**

### 2.2 ResNet50 的改进

由于 ResNet 可以重复叠加深度网络,因此可以提取更多的网络特征。本文提出的方法只采用 ResNet50 网络模型中的第 3 个 block 单元块之前的网络架构作为主干网络,第 3 个 block 单元块输出的 feature mapping 大小为 40×30×1 024。本文网络架构详见表 1。

**表 1 本文方法网络架构**

**Tab. 1 The network architecture of the method proposed in this paper**

| 序号 | 输入图片 | 网络层 | 输出(feature mapping) |
|---|---|---|---|
| 1 | 640×480×3 | 7×7 Conv, 64, stride 2 | 320×240×64 |
| 2 | 320×240×64 | 3×3 MaxPool, stride 2 | 160×120×64 |
| 3 | 160×120×64 | Residual Block 1 | 160×120×256 |
| 4 | 160×120×256 | Residual Block 2 | 80×60×512 |
| 5 | **80×60×512** | Residual Block 3 | 40×30×1 024 |
| 6 | 40×30×1 024 | 1×1 Conv, 128 | 40×30×128 |
| 7 | 40×30×1 024 | 3×3 Upconv, 512, factor 2 | **80×60×512** |
| 8 | **80×60×1 024** | 3×3 Conv, 512 | 80×60×512 |
| 9 | 80×60×512 | 3×3 Upconv, 256, factor 2 | **160×120×256** |
| 10 | **160×120×512** | 3×3 Conv, 256 | 160×120×256 |
| 11 | 160×120×256 | 1×1 Conv, 128 | 160×120×128 |

由表 1 可见,第 3 个 Block 单元块之后,在输出 40×30×1 024的 feature mapping 的基础上,依次进行 1×1 的卷积、上采样 3×3 卷积、3×3 卷积、上采样 3×3 卷积、3×3 卷积、1×1 卷积等操作,从而获得不同粒度的特征图。

通过附加的卷积层,获得了粗糙级特征图。精细级特征图是通过进一步的卷积层以及上采样和跳过连接获得的。粗略特征图和精细特征图的大小分别是原始图像的 1/16 和 1/4,且都具有 128 维的向量。精细级别的局部窗口 $W$ 的大小是精细级别特

征图大小的 1/8。

## 3　基于相机位姿的方法

如果只给出具有相机位姿的图像对，则不适用于标准深度度量学习方法。因此，本文设计了一种利用 IMU 数据进行特征描述子的学习方法。该方法中将相对相机位姿转换成图像对之间的对极约束，确保预测的匹配关系服从对极约束。考虑到该约束是实施在像素坐标上的，因此必须使对应的坐标相对于特征描述符是可微的，为了实现该目的，本文提出了可区分的匹配层方法。

### 3.1　损失函数

本文训练数据由含有位姿信息的图像对组成。为了利用这些数据训练特征描述子的关联匹配，使用 2 个互补的损失函数项：对极损失项和循环一致损失项。

图 1 中，$P_1$ 是查询点、$P_2$ 是预测的对应关系；对极损失函数 $L_{epipolar}$ 是 $P_2$ 和真实对极线 $FP_1$ 之间的距离；循环一致性损失 $L_{cycle}$ 是 $P_1$ 与其前后对应点（绿色）之间的 $L_2$ 距离。

给定一对图像 $I_1$ 和 $I_2$ 的相对位姿和相机内参，就可以计算基本矩阵 $\boldsymbol{F}$[8]。极线约束指出，如果 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是真实匹配，则 $\boldsymbol{p}_2^{\mathrm{T}}\boldsymbol{F}\boldsymbol{p}_1 = 0$ 成立，其中 $\boldsymbol{F}\boldsymbol{p}_1 = 0$ 可以解释为对应于 $I_2$ 中 $\boldsymbol{p}_1$ 的对极线。本文将 $\boldsymbol{p}_1$ 视为查询点，然后根据预测的对应位置与真实的对极线之间的距离，将此约束重新化为对极损失：

$$L_{epipolar}(\boldsymbol{p}_1) = distance(h_{1\to 2}(\boldsymbol{p}_1), \boldsymbol{F}\boldsymbol{p}_1) \tag{8}$$

其中，$h_{1\to 2}(\boldsymbol{p}_1)$ 是 $I_1$ 中点 $\boldsymbol{p}_1$ 在 $I_2$ 中的预测对应关系，而 $distance(\cdot,\cdot)$ 是点与线之间的距离。

单独的对极损失函数，预测的匹配点位于对极线上，而不是真实的对应匹配关系（该位置在该线上的未知位置）。为了提供额外的约束，本文还引入了循环一致性损失，确保该点在空间上接近其自身：

$$L_{cycle}(\boldsymbol{p}_1) = \| h_{2\to 1}(h_{1\to 2}(\boldsymbol{p}_1)) - \boldsymbol{p}_1 \|_2 \tag{9}$$

对于每个图像对，总目标是对极损失函数项和循环一致性损失项的加权总和，共计 $n$ 个采样查询点，可以表示为：

$$L_{all}(I_1, I_2) = \sum_{i=0}^{n} [L_{epipolar}(p_1^i) + \alpha L_{cycle}(p_1^i)] \tag{10}$$

其中，$p_1^i$ 是 $I_1$ 中的第 $i$ 个训练点，$\alpha$ 是周期一致性损失项的权重。

对极约束实际上为特征描述子学习提供了足够的监督。其关键原因是，对极约束抑制了许多不正确的对应关系。而且，在满足对极约束的所有有效预测中，鉴于其局部外观相似性，真正的对应关系最有可能具有相似的特征编码。因此，通过在所有训练数据上聚合这样的几何约束，使网络学会对真实对应之间的相似性进行编码，从而产生有效的特征描述子。

尽管本文的重点是仅从相机的位姿中学习，但当有真实匹配关系可用时，也可以使用真实的对应关系进行训练。在这种情况下，可以将损失函数替换为预测和真实对应关系的像素位置之间的 $L_2$ 距离。通过真实匹配关系训练的方法，比通过照相机姿势训练的方法会获得更好的性能，两者均优于先前的完全监督方法。

### 3.2　可区分的匹配层

损失函数是预测对应像素位置的函数，但若使用梯度下降法，则像素位置相对于网络参数是可区分的。传统方法是通过识别最近邻匹配来建立对应关系，但这是一种不可微分的操作。

针对上述问题，本文提出了可区分的匹配层方法。对于给定的图像对，首先使用卷积神经网络提取特征描述子 $M_1$、$M_2$，为了计算 $I_1$ 中查询点 $x_1$ 的对应关系，将 $x_1$ 处的特征描述子（用 $M_1(x_1)$ 表示）与 $M_2$ 相关联。接下来进行 2D softmax 操作，可以得到 $x_1$ 在 $I_2$ 中的二维概率[9]分布 $p(x \mid x_1, \boldsymbol{M}_1, \boldsymbol{M}_2)$：

$$p(x \mid x_1, \boldsymbol{M}_1, \boldsymbol{M}_2) = \frac{\exp(\boldsymbol{M}_1(x_1)^{\mathrm{T}} \boldsymbol{M}_2(x))}{\sum_{y \in I_2} \exp(\boldsymbol{M}_1(x_1)^{\mathrm{T}} \boldsymbol{M}_2(y))} \tag{11}$$

其中，变量 $y$ 表示在 $I_2$ 的像素坐标上变化。

计算单个 2D 匹配作为此分布的期望：

$$\hat{x}_2 = h_{1\to 2}(x_1) = \sum_{x \in I_2} x \cdot p(x \mid x_1, \boldsymbol{M}_1, \boldsymbol{M}_2) \tag{12}$$

使用可区分的匹配层，使整个网络系统可以进行端到端训练。由于对应位置是根据特征描述子之间相关性计算的，因此将有助于特征描述子的训练学习。

## 4　实验结果分析

### 4.1　数据集

本文提出的方法使用 MegaDepth 数据集[10]对网络进行训练。该数据集由 196 个不同场景组成。其中 130 个场景用于训练，其余场景用于验证和测试。数据集提供了数百万个具有已知相机位姿的图

像训练匹配对,在此仅使用所提供的相机位姿和相机内参在这些图像匹配对上进行训练。

### 4.2 训练过程

本文使用 Adam 训练网络,其基本学习率为 $1\times10^{-4}$,循环一致性项的权重设置为 0.1。由于内存所限,在每个训练图像对中使用 400 个查询点。这些查询点由 80%SIFT 关键点和 20%随机点组成。

训练所使用的设备为:RTX3060 GPU,Ubuntu 18.04 操作系统与 Intel i7 第十代 CPU。

### 4.3 实验结果

本文在 MegaDepth 数据集上对 POSE 特征描述子进行测试。给定一对图像后,在 2 个图像中提取关键点,并使用特征描述子进行描述。对每个图像匹配对之间的匹配数进行统计(仅考虑最近邻匹配),并将 SIFT、LF-Net[11] 与 POSE 测试结果进行对比。

去除异常点之后的匹配结果如图 4~图 6 所示。其中,图 4 的 SIFT 特征描述子代表传统人工设计方法;图 5 代表 LF-Net 深度学习方法;图 6 为采用本文提出的方法。根据对比的结果可以看出,本文方法得到的结果使精度得到了提高。

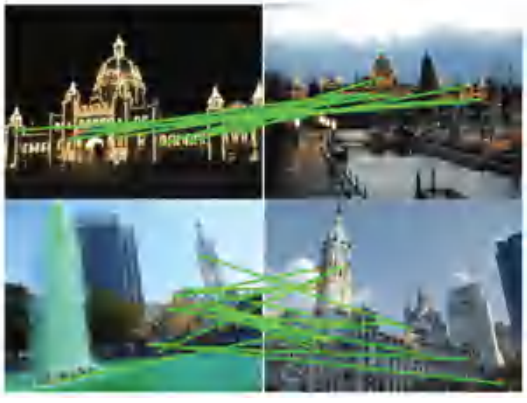

**图 4 SIFT 描述子**

**Fig. 4 SIFT descriptor**

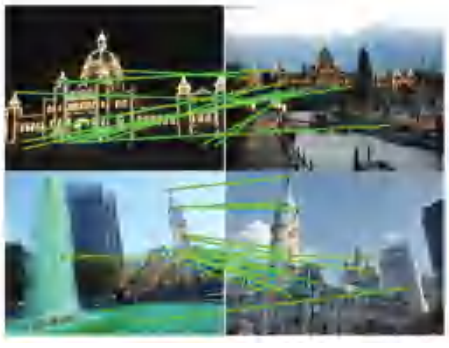

**图 5 LF-Net 描述子**

**Fig. 5 LF-Net descriptor**

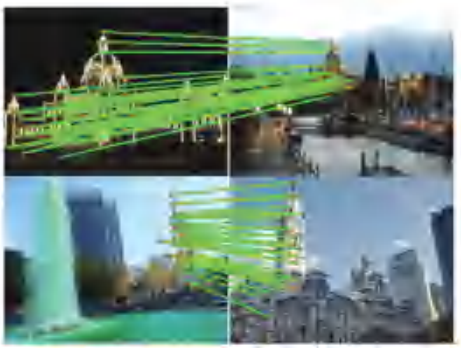

**图 6 POSE 描述子**

**Fig. 6 POSE descriptor**

实验不仅给出可视化的结果,并对几种方法进行了量化处理。3 种特征描述子对比结果见表 2。

**表 2 POSE 特征描述子与其它模型的对比**

**Tab. 2 Comparison of POSE feature descriptor with other models**

| 方法 | 特征 | 匹配 |
|---|---|---|
| SIFT | 440 | 120 |
| LF-Net | 170 | 80 |
| POSE | 470 | 190 |

从表 2 中可以看出,POSE 特征描述子不仅可以从图像中提取到更多的特征点,同时也提高了特征描述子的匹配度。

## 5 结束语

文中提出了一种新颖的特征描述子学习框架,该框架仅使用 IMU 获取的相机位姿监督进行训练,利用对极几何约束构造损失函数。实验表明,在不使用任何特征点对应匹配关系进行训练的情况下,其性能优于受到完全监督的其它算法。接下来的工作中,将进一步研究如何提高学习到的特征描述子对图像旋转的不变性、基于位姿监督的方法和传统度量学习方法是否有相互补充的可能性,以及相应的组合是否可以带来更好的性能。

## 参考文献

[1] SCHONBERGER J L, FRAHM J M. Structure-from-motion revisited[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: 2016, 2016: 4104-4113.

[2] SATTLER T, MADDERN W, TOFT C, et al. Benchmarking 6dof outdoor visual localization in changing conditions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 8601-8610.

[3] MISHCHUK A, MISHKIN D, RADENOVIC F, et al. Working hard to know your neighbor's margins: Local descriptor learning loss[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Long Beach California USA: NIPS, 2017: 4829-4840.

(下转第 207 页)