激光与光电子学进展

面向细节保持的特征描述子提取算法

龙涛，苏畅，王建*

天津大学电气自动化与信息工程学院，天津 300072

摘要 检测图像中的显著关键点并提取特征描述子是视觉里程计和同步定位与建图系统等计算机视觉任务中的重要环节。特征点提取算法的主要目标是检测准确的关键点位置并提取可靠的特征描述子。可靠的特征描述子应对旋转、尺度缩放、光照变化、视角变化、噪声等保持一定程度的稳定性。目前基于深度学习的方法由于描述子特征在下采样过程中存在图像信息丢失，导致描述子可靠性和特征匹配准确度降低。针对这一问题，提出了一种面向细节保持的特征描述子提取网络。该网络融合浅层细节特征和深层语义特征，将描述子特征上采样到更高的空间分辨率，并结合注意力机制，使用局部特征（角点、线段、纹理等）、语义特征和全局特征来改进特征点检测，提高特征描述子可靠性。在Hpatches数据集上的实验结果表明，所提方法的匹配准确度为55.5%。输入图像分辨率为480×640时，所提方法的单应性估计准确度比现有方法高5.9个百分点。实验结果表明了所提方法的有效性。

关键词 机器视觉；特征点检测；深度学习；卷积神经网络

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.2215002

Learning Feature Point Descriptors for Detail Preservation

Long Tao, Su Chang, Wang Jian*

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Detecting salient key points in images and extracting feature descriptors are important components of computer vision tasks such as visual odometry and simultaneous localization and mapping systems. The main goal of the feature point extraction algorithm is to detect accurate key point positions and extract reliable feature descriptors. Reliable feature descriptors should maintain stability against rotation, scale scaling, illumination changes, viewing angle changes, noise, etc. Due to the loss of image information during the downsampling process in recent deep learning-based feature point extraction algorithms, the reliability of the descriptor and accuracy of feature matching are reduced. This study proposes a network structure to detect detail-preserving oriented feature descriptors to solve this problem. The proposed network fuses shallow detail and deep semantic features to sample the descriptors to a higher resolution. Combined with the attention mechanism, local (corners, lines, textures, etc.), semantic, and global features are used to improve the detection of feature points and the reliability of feature descriptors. Experiments on the Hpatches dataset show that the matching accuracy of the proposed method is 55.5%. Additionally, when the input image resolution is 480×640, the homography estimation accuracy of the proposed method is 5.9 percentage points higher than that of the existing method. These results demonstrate the effectiveness of the proposed method.

Key words machine vision; feature point detection; deep learning; convolutional neural network

1 引 言

特征点由关键点和描述子两部分组成。关键点表示特征点在图像中的位置，描述子表示该点的局部特征。检测图像的特征点并在视图间匹配是许多机器人系统的前端任务，例如同步定位与建图（SLAM）[1]、三维重建（SfM）[2]和视觉里程计[3]，这些任务需要检测出对光照效果、视点变化和尺度变化等保持不变的显著点，以保证其在序列图像中能被重识别。然而，这些任务仍然主要依赖手工设计的图像特征，例如特征点提取算法（SIFT[4]、SURF[5]、ORB[6]、BRISK[7]）。

深度学习方法已经大大改进了许多计算机视觉应用，包括目标检测，语义分割等。大多数基于深度学习算法需要监督即依赖标签，人工标注往往需要付出昂

收稿日期：2021-08-19；**修回日期**：2021-09-04；**录用日期**：2021-09-24

基金项目：国家自然科学基金（61632018）

通信作者：*jianwang@tju.edu.cn

贵的代价。现实中很难为监督特征点检测去标注数据，因为标注者不能很容易地识别图像中的那些能被重新识别的显著区域，并且难以保证选取标准的一致性。因此，对于特征点检测来说，不依赖于真值标签的训练非常重要。

目前有许多基于深度学习的特征点提取算法，并且在性能上已经超过传统的特征点提取算法[8]。Learned invariant feature transform（LIFT）[9]使用3个模块预测特征点和描述子：1个用于创建特征点分数图的检测器、1个用于预测图像块方向的方向估计器和1个描述子模块。空间变换网络（STN）[10]通过估计的方向旋转每个图像块，然后把旋转后的图像块输入描述子网络。LIFT在单个图像块上是端到端可微的，但该模型不能对整张图像进行训练。该方法是多阶段的，并且需要SfM模块来指导训练。此外，LIFT框架中的每个模块都不共享计算。LF-Net[11]类似于LIFT，与LIFT不同的是，位置、旋转和尺度由单个模块估计。LF-Net能够从零开始对完整图像进行训练，速度很快，该模型在SfM图像匹配中具有最好的性能。但该框架在训练期间需要SfM模块的输出参与，检测器和描述子模块之间共享计算，基于图像块的方法也限制了网络能够学习描述子的区域。SuperPoint[12]也能够预测关键点位置和描述符，其关键点检测器和描述子模块共享大部分计算，因此速度很快。该方法先在基于几何形状的合成数据集上训练，然后在真实数据集上用孪生网络进行训练。该方法的缺点在于特征点是由人工定义的几何形状的角点。UnsuperPoint[13]是一种基于深度学习的端到端特征点提取算法，它只需要以自监督的方式进行单阶段训练。该方法采用简单的单应性变换及非空间图像增强来创建二维合成图像，通过原图和合成图像来训练自监督关键点估计模型。

UnsuperPoint方法[13]采用简单的VGG网络结构，下采样后的深层特征经过归一化之后即为描述子向量。对于一些特殊场景，例如特征点距离相机很远，特征点附近的局部图像尺度很小，因此该点局部细节信息在下采样的过程存在丢失现象，导致描述子无法匹配或误匹配，降低了描述子匹配分数和单应性估计准确度。

本文基于UnsuperPoint方法，针对上述问题提出了一种融合不同尺度低层特征的网络结构，提高描述子特征图的分辨率，确保描述子信息的完整性、丰富性和准确性，提升描述子预测网络对尺度缩放的不变性。还引入通道注意力机制——压缩激励[14]模块，来调整特征融合后描述子特征图的通道分布。原因在于描述子特征融合了浅层细节特征（角点、线段、纹理、颜色）和深层高级特征，这些信息分布在不同的维度上，对于不同的场景，每种特征的重要性不同，通道注意力机制可以让网络学习各个通道的重要程度，来增强该通道对应的特征，最终提高描述子的可区分度，减少误匹配。实验结果表明，所提方法在Hpatches数据集[8]评价基准下取得了有效提升。

2　原理和方法

2.1　特征点提取网络结构

整个网络结构由共享主干网络和3个具有特定任务的子模块构成，如图1所示。主干网络将彩色图像

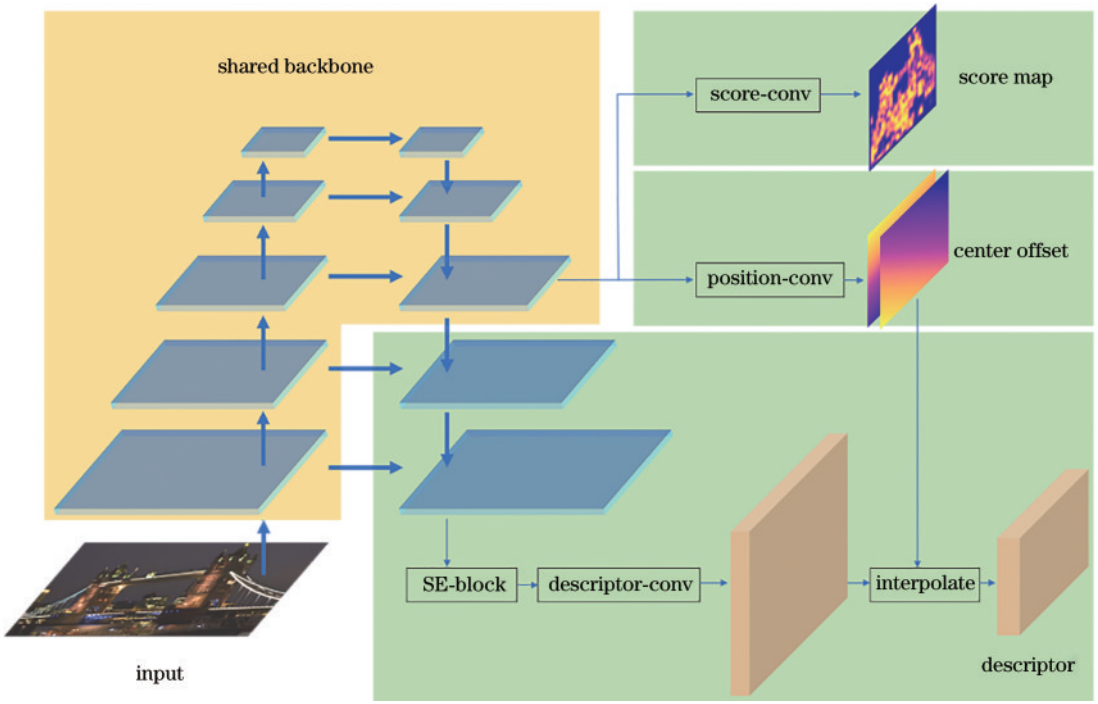

图1　特征点提取神经网络结构

Fig. 1　Neural network architecture of feature point extraction

作为输入，并提供一个降采样后的特征图供后续任务的子模块处理。

后续的3个子模块分别是分数预测模块、位置预测模块和描述子预测模块。与UnsuperPoint的网络结构差异主要为主干网络和描述子模块的不同。在不显著增加计算量的情况下，所提方法较基础结构有所提升。整个网络的卷积结构模型可以处理任意大小的输入图像。网络的组合输出类似于传统的特征点提取算法的输出，输出为一个点的分数、位置和描述子，因此该网络可以作为传统的基于特征点的系统（例如SLAM）前端的替代方案。

2.2　主干网络

主干网络的前五层为Resnet-18结构[15]，可以得到尺寸逐渐减半、通道数为64-64-128-256-512的特征图。采用一种类似于U-Net[16]的形式，特征图先通过卷积核大小为3、步长为1的卷积层将通道数降低为256-256-128-64-32，然后经过最近邻插值法上采样，将上采样后的特征图和Resnet输出中尺寸一致的特征图拼接起来，将拼接后的特征通过卷积核大小为3、步长为1、通道数为256-256-128-64的卷积层，最终得到信息融合后的分辨率更高的特征图。每个卷积层都含有批标准化层和带泄露修正线性单元（leaky ReLU）激活函数。若输入图像大小为$H\times W$，将大小为$H/8\times W/8$、通道数为256的特征图作为分数预测分支和位置预测分支的输入，这种做法与UnsuperPoint是一致的。不同的是，本实验组将分辨率更高、通道数更少、低层局部信息更丰富的特征图作为描述子预测分支的输入。

2.3　分数预测模块

分数预测模块为大小为$H/8\times W/8$、通道数为256的特征图中每一个点回归得到一个分数值，该点对应原图上大小为8×8的区域，该分数值用于判断该区域是否含有特征点。分数预测模块包含两个通道数为256和1的卷积层，sigmoid激活函数保证分数值位于区间[0,1]。最终选择分数最高的N个点作为特征点。

2.4　位置预测模块

位置预测模块为特征图中每个点回归得到相对偏移位置（X轴方向和Y轴方向相对中心偏移度），并将其映射到图像像素坐标。该模块包含两个卷积层，通道数分别是256和2，最后一层之后是tanh激活函数，保证位置偏移数值位于区间[−1,1]。图2以大小为24×24的输入图像为例，演示了特征点位置预测过程，所提方法对每个8×8区域中只选取一个特征点，所以只会预测9个特征点。

由相对中心偏移$P_{\text{center-offset}}$到图像像素坐标P_{map}的映射可描述为

$$P_{\text{map},x}(r,c)=\left[c+P_{\text{center-offset},x}(r,c)+1\right]\frac{f_{\text{downsample}}}{2},\quad(1)$$

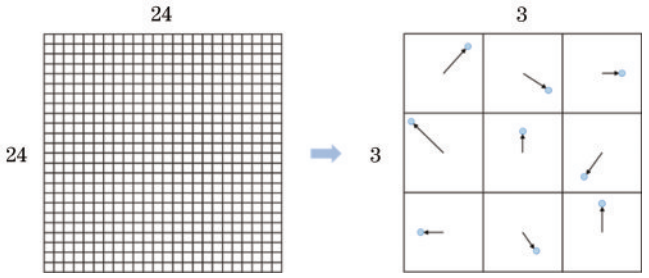

图2　位置预测中心偏移示意图

Fig. 2　Schematic diagram of center offset of position prediction

$$P_{\text{map},y}(r,c)=\left[c+P_{\text{center-offset},y}(r,c)+1\right]\frac{f_{\text{downsample}}}{2},\quad(2)$$

式中：r、c为特征图的像素坐标索引；特征图的下采样倍数$f_{\text{downsample}}=8$。需要注意的是，由于在每个8×8区域上只选取一个特征点，这种方法有非极大抑制（NMS）的作用。

2.5　描述子预测模块

描述子预测模块为输入特征图中的每一点生成一个描述子向量。输入的特征图首先通过降维、上采样操作，然后和同尺寸的低层特征拼接在一起，再经过一次卷积，得到融合细节局部信息的特征图。经过两次上述操作，特征图的尺寸提升到$H/2\times W/2$。高分辨率的特征图经过压缩激励模块，根据对各通道的依赖程度进行调整，调整后的特征图需要经过两次卷积核为3×3、步长为1、通道数为64的卷积层，最后一层不含有激活函数。根据位置预测模块得到的坐标使用插值法得到更精确的描述子特征图，最后经过L2正则归一化得到描述子向量。

2.6　压缩激励模块

通过学习的方式来自动获取到每个特征通道的重要程度，然后依照这个重要程度去增强局部特征（颜色、角点、线段、纹理）和语义特征并抑制对当前任务用处不大的噪声和冗余信息。本实验组使用全局平均池化作为压缩操作来压缩空间维度上的特征，将每个二维的特征通道变成一个实数，获得通道特征上相应的全局分布。先经过一个全连接层把通道数降低为输入的1/16，经过激活函数（ReLU）后再经过一个全连接层恢复到原来的通道数。然后通过sigmoid函数获得归一化权重，最后经过scale操作将归一化之后的权重加权到每个通道特征上。

2.7　自监督训练框架

和UnsuperPoint中自监督方法相同，本实验组也使用孪生网络来进行自监督训练。原图像$\boldsymbol{I}^{\text{s}}$通过随机单应性变换$T$（旋转、缩放、倾斜和透视变换）进行空间变换得到$\boldsymbol{I}^{\text{w}}$，然后进行独立的随机非空间图像增强（如亮度和噪声）变换。将增强后的图像对分别送入神经网络预测特征点，记为$\boldsymbol{I}\rightarrow\{\boldsymbol{s},\boldsymbol{p},\boldsymbol{f}\}$，其中$\boldsymbol{s}$是特征点分数图，$\boldsymbol{p}$为特征点的图像像素坐标，$\boldsymbol{f}$是特征点描述子

向量的集合。原图上预测的特征点通过单应性 T 变换与变换后图像上的特征点在空间位置上对齐，并取欧氏距离最小的点对为匹配点对。在这些匹配点对之间引入损失函数来训练模型。

2.8 损失函数

如何构建损失函数对于神经网络的训练来说十分重要，通过梯度下降法降低损失，优化网络参数。本实验组用来训练网络的损失函数可描述为

$$\mathcal{L}=L_{\mathrm{loc}}+\alpha L_{\mathrm{score}}+\beta L_{\mathrm{score\text{-}loc}}+\gamma L_{\mathrm{desc}}, \tag{3}$$

式中：L_{loc} 为位置损失；L_{score} 为分数损失，超参数 α 为分数损失的权重；$L_{\mathrm{score\text{-}loc}}$ 为位置-分数联合损失，β 为位置-分数联合损失的权重；L_{desc} 为描述子损失，γ 为描述子损失的权重。为了更好说明各项损失函数的意义，需要先从两个输入图像之间定义匹配点对集合 $\boldsymbol{M}$。在训练阶段输入一对图像 $\boldsymbol{I}^{\mathrm{s}}$ 和 $\boldsymbol{I}^{\mathrm{w}}$，得到两组特征点 $\{\boldsymbol{s}^{\mathrm{s}},\boldsymbol{p}^{\mathrm{s}},\boldsymbol{f}^{\mathrm{s}}\}$ 和 $\{\boldsymbol{s}^{\mathrm{w}},\boldsymbol{p}^{\mathrm{w}},\boldsymbol{f}^{\mathrm{w}}\}$。单应性变换 T 已知，可以把原图上特征点变换到另一张图像上，位置变化记为 $T\boldsymbol{p}^{\mathrm{s}}$ 或 $\boldsymbol{p}^{\mathrm{s}\to\mathrm{w}}$。对于每一个变化过来的特征点 $\boldsymbol{p}_i^{\mathrm{s}\to\mathrm{w}}$，$i$ 为图像中特征点索引，$i=1,2,\cdots,n$，基于欧氏距离选取 $\boldsymbol{p}^{\mathrm{w}}$ 中最近的点作为匹配点对。所有欧氏距离小于预先设定阈值 ε_{d} 的匹配点的集合即为 $\boldsymbol{M}$。假设集合 $\boldsymbol{M}$ 共有 K 个匹配点对，匹配点对集合 $\boldsymbol{M}$ 中任意点对的位置距离 d_k 的表达式为

$$d_k=\left\| T\boldsymbol{p}_k^{\mathrm{s}}-\boldsymbol{p}_k^{\mathrm{w}}\right\|=\left\| \boldsymbol{p}_k^{\mathrm{s}\to\mathrm{w}}-\boldsymbol{p}_k^{\mathrm{w}}\right\|, \tag{4}$$

式中：k 为匹配点对的索引，$k=1,2,\cdots,K$。

式(3)前三项损失继承于UnsuperPoint方法，是为了提高特征点检测器的可重复性，即无论摄像机视角如何变化，检测器总能检测到相同的特征点。换言之，检测器应该从多个摄像机视角预测捕捉场景中相同三维(3D)点的图像位置。

第1项损失的表达式为

$$L_{\mathrm{loc}}=\sum_{k=1}^{K}d_k。 \tag{5}$$

训练孪生网络的初始阶段，预测的特征点的位置是随机的，随着迭代次数的增加，损失将逐渐减少，匹配点对的位置距离逐渐减小，位置预测的精度将逐渐增加。第2项损失的表达式为

$$L_{\mathrm{score}}=\sum_{k=1}^{K}\left(s_k^{\mathrm{s}}-s_k^{\mathrm{w}}\right)^2。 \tag{6}$$

该项通过减少匹配点对分数的差异以确保匹配点对分数预测的相似性。第3项损失的目的在于确保预测分数代表特征点的置信度，分数最高的点应该是可重复性最强的点，分数最低的点应该是可重复性最差的点。匹配点对之间构建的第3项损失为

$$L_{\mathrm{score\text{-}loc}}=\sum_{k=1}^{K}\frac{s_k^{\mathrm{s}}+s_k^{\mathrm{w}}}{2}\left(d_k-\bar{d}\right), \tag{7}$$

$$\bar{d}=\frac{1}{K}\sum_{k=1}^{K}d_k。 \tag{8}$$

一个好的特征点为点对对应距离 d_k 较低的点。相反，对于一个不好的特征点，d_k 很大，因为网络无法一致地预测点的位置。对于式(7)：若匹配点对的像素距离 $d_k<\bar{d}$，网络模型必须学习预测高分以最小化损失；反之，对于 $d_k>\bar{d}$，网络模型必须学习预测低分以最小化损失。

本实验组采用文献[17]度量学习中的基于难样本挖掘的三元组损失来训练描述子分支。原图像中每一个关键点 $\boldsymbol{p}_i^{\mathrm{s}}$ 对应一个描述子 $\boldsymbol{f}_i^{\mathrm{s}}$，$\boldsymbol{f}_i^{\mathrm{s}}$ 是根据 $\boldsymbol{p}_i^{\mathrm{s}}$ 在高分辨率描述子特征图上采样得到的，把 $\boldsymbol{f}_i^{\mathrm{s}}$ 作为锚点，通过投影过来的关键点 $\boldsymbol{p}_i^{\mathrm{s}\to\mathrm{w}}$，在图像 $\boldsymbol{I}^{\mathrm{w}}$ 的高分辨率描述子特征图上采样得到的描述子作为正样本，记为 $\boldsymbol{f}_{i,+}^{\mathrm{w}}$。从图像 $\boldsymbol{I}^{\mathrm{w}}$ 的一组描述子 $\boldsymbol{f}^{\mathrm{w}}$ 中选取与锚点 $\boldsymbol{f}_i^{\mathrm{s}}$ 最相似的且与正样本不同的描述子向量作为负样本，记为 $\boldsymbol{f}_{i,-}^{\mathrm{w}}$。需要注意的是，这里的不同是指负样本 $\boldsymbol{f}_{i,-}^{\mathrm{w}}$ 对应的关键点坐标 $\boldsymbol{p}_i^{\mathrm{w}}$ 和投影坐标 $\boldsymbol{p}_i^{\mathrm{s}\to\mathrm{w}}$ 在每个维度的坐标差大于设定阈值 $\varepsilon_{\mathrm{relax}}$。因此三元组损失的表达式为

$$L_{\mathrm{desc}}=\sum_{i}^{S}\max\left(0,\left\| \boldsymbol{f}_i^{\mathrm{s}},\boldsymbol{f}_{i,+}^{\mathrm{w}}\right\|-\left\| \boldsymbol{f}_i^{\mathrm{s}},\boldsymbol{f}_{i,-}^{\mathrm{w}}\right\|+m\right),S=\frac{H}{8}\times\frac{W}{8}, \tag{9}$$

式中：m 为距离参数，表示不相似的描述子向量在描述子空间中的距离限制。在训练过程中，该损失能逐渐减小锚点和正样本之间的距离，并使锚点和负样本之间的距离增加。除了匹配点之外的任何描述子样本都可以用锚点的负样本，但是选择最难的负样本对损失函数的贡献最大，可以加速度量学习。第4项损失函数的意义在于训练描述子分支，使其满足描述子应具有可区分性的要求。

3 实验结果与分析

3.1 实验细节

本实验组使用Pytorch训练神经网络，选择MS COCO数据集[18]中的118287张训练集图像作为本实验的训练集，训练过程无需使用标签，输入图像尺寸为 256×320。实验使用ADAM梯度下降算法优化网络参数，学习率设置为 10^{-3}，batch size设置为8，总共训练50个epoch，网络权重随机初始化。损失权重超参数设置如下：$\alpha=1$、$\beta=1$、$\gamma=2$。阈值参数 $\varepsilon_{\mathrm{d}}=4$，描述子损失中参数 $\varepsilon_{\mathrm{relax}}=8$，三元组损失参数 $m=0.2$。

在孪生网络训练结构中，单应性变换 T 操作包括裁剪、缩放、旋转和透视变换。首先将图像裁剪为原始图像分辨率的0.7，其他变换数值从预定范围中随机选取，尺度变换范围是[0.8,1.2]，旋转角度范围是[0°,90°]，透视变换范围是[0,0.2]。

3.2 评价指标

实验使用SuperPoint的评估指标[12]，分别是可重复率(RR)、定位误差(LE)、匹配分数(MS)和单应性估计准确度(HA)。通过可重复率和定位误差评估关

键点位置检测器，在单应性估计框架下通过测量匹配分数和单应估计准确度来评估整个检测器（分数、位置和描述子）。

可重复率衡量关键点的质量，可重复率是两个视角观察到的点数与总点数之间的比率[19]。对于平面场景，可以通过简单地使用单应性矩阵将点从一个视图映射到另一个视图来建立两个摄像机视图之间的特征点对应关系。关联点对距离低于某个像素距离（$\rho=3$）时为正确关联点，可重复率即为正确关联的比率。定位误差为所有映射点和其关联点之间的平局像素距离。

单应性估计过程如图3所示。先使用神经网络从同一平面场景的两个图像中提取特征点，并从每幅图像中只选取分数最高的N个点作为最终的特征点。使用最近邻（暴力）匹配方法匹配两幅图像的描述子，调用OpenCV工具库，使用RANSAC算法[20]估计单应性矩阵，滤除误匹配得到最终的匹配集。调用算法的参数如下：最大迭代次数为5000、置信阈值为0.9995、错误阈值为3。

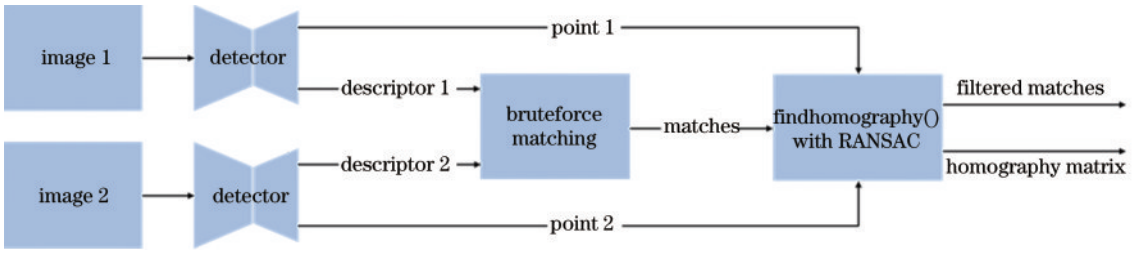

图3　单应性估计流程图

Fig. 3　Flow chart of homography estimation

匹配分数是正确匹配与共享视图中所有点之间的比率。此处正确的匹配是对于采用最近邻（暴力）匹配两幅图像的描述子得到的匹配点对来说的，将第1幅图像上的点通过单应性矩阵真值变换到第2幅图像上，若映射点和对应点的坐标像素距离小于某个像素距离（$\rho=3$），则判定为正确匹配。

单应性估计准确度是正确估计的单应性矩阵与总数的比率，如图4所示。首先利用单应性估计矩阵（$\boldsymbol{H}_{\mathrm{est}}$）和单应性真值矩阵（$\boldsymbol{H}_{\mathrm{gt}}$）对图像边框分别进行变换操作，变换后边框顶点之间的像素距离即为单应性估计误差，如图4中虚线所示。为了量化估计的单应性是否正确，若单应性估计误差小于预设误差阈值ε则为正确估计。本实验在多个预设值下统计单应性估计准确度，即分别统计$\varepsilon=1$、3、5时的单应性估计准确度。

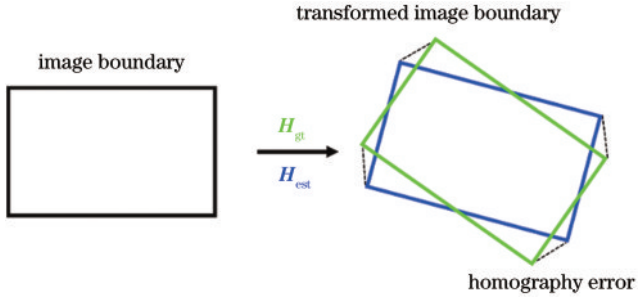

图4　单应性误差计算原理图

Fig. 4　Schematic diagram of homography error calculation

3.3　修改网络结构对性能的影响实验

本实验在Hpatches数据集的完整图像序列上评估方法的性能。该数据集包含57个光照变化组和59个视角变化组。每组包含平面场景的6张图像（1个参考图像和5个目标图像）和将参考图像映射到5个目标图像的单应性变换矩阵真值，光照变化组的6张图像是从同一视角不同光照条件下拍摄同一平面场景得到的，视角变化组的6张图像是相同光照条件下从不同视角拍摄同一场景获得的。将参考图像与每一个目标图像作为评价算法的输入，总共有580（$57\times5+59\times5$）个图像对，每个图像对都会计算相应的评价指标，最终的度量结果是对所有图像对的评价指标取均值得到的，如表1所示，其中粗体表示最优结果。

表1展示了不同的网络结构对性能的影响。验证过程中输入图像的分辨率为256×320，选择分数最高的前300个特征点。Baseline的主干网络是VGG网络，深层特征输入描述子分支得到通道数为256的语义描述子。V1把Baseline中的主干网络替换成残差网络结构Resnet，实验结果表明，此改进大大提升了特征点位置回归的准确度和匹配分数，原因是残差网络在表征性能上更好，但是单应性估计准确度并没有太大提升，这是因为更换主干网络并没有解决细节信息丢失的问题。V2在V1的基础上融合深层语义特征和浅层局部特征，将融合后的特征送入描述子分支，得到

表1　不同网络结构的实验结果对比

Table 1　Comparison of experimental results of different network structures

Method	Repeat	LE	HA-1	HA-3	HA-5	MS
Baseline	0.633	1.044	0.503	0.796	0.868	0.491
V1	0.675	**0.831**	0.505	0.822	0.897	**0.576**
V2	**0.676**	0.856	0.581	0.866	0.903	0.554
V3	0.669	0.842	**0.586**	**0.871**	**0.912**	0.555

信息更加全面的描述子向量，且描述子的维度降为64。实验结果显示，单应性估计准确度显著提升，验证了所提方法的有效性。图5详细展示了Baseline和V1方法中存在的问题，在一些视角变化较大、尺度缩放程度大的情况中，右图方框在下采样过程丢失了细节特征，根据红点位置插值得到的描述子信息不准确，导致左右两图里红点的描述子无法匹配。V2由于融合了低层细节特征，并把描述子特征图分辨率提升到图像分辨率的一半，最大程度上保留了细节信息，确保了该显著点的描述子是准确可靠的，进一步验证了所提方法的正确性和有效性。V2相比V1匹配分数略微降低，原因在于为了保证神经网络整体参数量大体不变，描述子维度由256降低为64。尽管匹配分数略微减少，考虑到单应性估计准确的显著提高，仍然可以说明所提方法是有效的。V3在V2的基础上加入压缩激励模块，引入全局信息调节通道特征分布。实验结果表明，单应性估计准确度得到小幅度提升，验证了通道注意力机制在这个任务中是有效的。

图5 基础方法特征匹配失败案例

Fig. 5 Failure case of baseline feature matching

3.4 实验结果对比

为了验证所提方法的优越性，在HPatches数据集上将所提方法和其他先进方法及传统方法进行了比较。实验分别在低分辨率和高分辨率的输入图像上测试，低分辨率图像保留300个特征点，高分辨率图像保留1000个特征点，高分辨率图像尺寸为480×640。关键点检测指标的实验结果如表2所示。对于可重复率指标，所提方法明显优于其他方法。对于定位误差，UnsuperPoint在低分辨率图像作为输入时表现更好。原因是UnsuperPoint方法中损失函数采用关键点位置均匀分布正则项，即要求预测的关键点位置在8×8的区域内均匀分布，所以定位误差总体比所提方法好。此外，所提方法在高分辨率图像作为输入时优于其他方法。

表2 不同方法的关键点检测性能比较

Table 2 Comparison of key point detection performance of different methods

Method	Repeatability rate		Localization error	
	Low resolution	High resolution	Low resolution	High resolution
ORB	0.532	0.525	1.429	1.430
SURF	0.491	0.468	1.150	1.244
BRISK	0.566	0.505	1.077	1.207
SIFT	0.451	0.421	0.855	1.011
LF-Net(indoor)	0.486	0.467	1.341	1.385
LF-Net(outdoor)	0.538	0.523	1.084	1.183
SuperPoint	0.631	0.593	1.109	1.212
UnsuperPoint	0.645	0.612	**0.832**	0.991
Proposed method	**0.669**	**0.663**	0.842	**0.926**

单应性估计和匹配性能实验结果如表3所示。与基于传统特征点提取算法相比，自监督学习的方法在HA-3、HA-5和匹配分数上性能更好。但是对于更加严格阈值下的单应性估计HA-1，传统方法SIFT的精度更高。这是因为在测试集HPatches中，存在一部分旋转变化幅度较大的图像对。传统的SIFT方法对角度具有良好的旋转不变性，单应性估计误差大概率小于1；由于卷积神经网络的数学性质，其在旋转不变性上存在先天劣势，单应性估计误差大概率大于1。SIFT的性能在旋转场景下远远超过基于深度学习的方法，优势甚至盖过了其他场景的劣势，所以数据上HA-1比较高。但在HA-3、HA-5标准下，基于深度学习算法的优势逐渐追赶上来，总体好于SIFT方法。

表3 不同方法单应性估计和匹配性能比较

Table 3 Comparison of homography estimation and matching performance of different methods

Method	Low resolution, 300 points				High resolution, 1000 points			
	HA-1	HA-3	HA-5	MS	HA-1	HA-3	HA-5	MS
ORB	0.131	0.422	0.540	0.218	0.286	0.607	0.71	0.204
SURF	0.397	0.702	0.762	0.255	0.421	0.745	0.812	0.230
BRISK	0.414	0.767	0.826	0.258	0.300	0.653	0.746	0.211
SIFT	**0.622**	0.845	0.878	0.304	**0.602**	0.833	0.876	0.265
LF-Net(indoor)	0.183	0.628	0.779	0.326	0.231	0.679	0.803	0.287
LF-Net(outdoor)	0.347	0.728	0.831	0.296	0.400	0.745	0.834	0.241
SuperPoint	0.491	0.833	0.893	0.318	0.509	0.834	0.900	0.281
UnsuperPoint	0.579	0.855	0.903	0.424	0.493	**0.843**	0.905	0.383
Proposed method	0.586	**0.871**	**0.912**	**0.555**	0.552	0.840	**0.916**	**0.508**

所提方法相比于其他基于学习的方法表现更好，即无论是低分辨率输入还是高分辨率输入，在不同阈值下的单应性估计准确度和匹配分数更高。特别是在高分辨率输入下，所提方法 HA-1 准确度明显高于 UnsuperPoint 方法。

还在 HPatches 数据集的两个子集——光照变化组和视角变化组上分别测试各项指标，并与目前先进的基于学习的方法[14]进行了比较，实验结果如表 4 所示。从表 4 可以发现，在光照变化组上估计单应性准确度明显高于视角变化组，这是由于视角变化组中含有旋转幅度较大的图像对，全卷积神经网络无法处理这类极端问题，因此效果较差。所提方法在更具有挑战性的视角变化组上的性能略优于文献[14]中的方法，即所提多尺度特征融合和通道注意力机制是有效的。图 6 展示了所提方法在 HPatches 数据集上特征描述子成功匹配的可视化结果，在比较有挑战的图像对上（光照改变剧烈、旋转角度较大和视角变化幅度较大）可以取得很好的特征匹配结果。

表 4　不同数据子集上实验结果对比

Table 4　Comparison of experimental results on different data subsets

Method	Hpatches subset	Repeat	LE	HA-1	HA-3	HA-5	MS
Outlier_rejection[14]	ALL	0.686	0.890	0.595	0.871	0.912	0.544
	Illumination	0.678	0.826	0.753	0.942	0.984	0.614
	Viewpoint	0.693	0.953	0.494	0.801	0.857	0.479
Proposed method	ALL	0.669	0.842	0.586	0.871	0.912	0.555
	Illumination	0.643	0.789	0.642	0.933	0.965	0.576
	Viewpoint	0.695	0.893	0.532	0.810	0.861	0.534

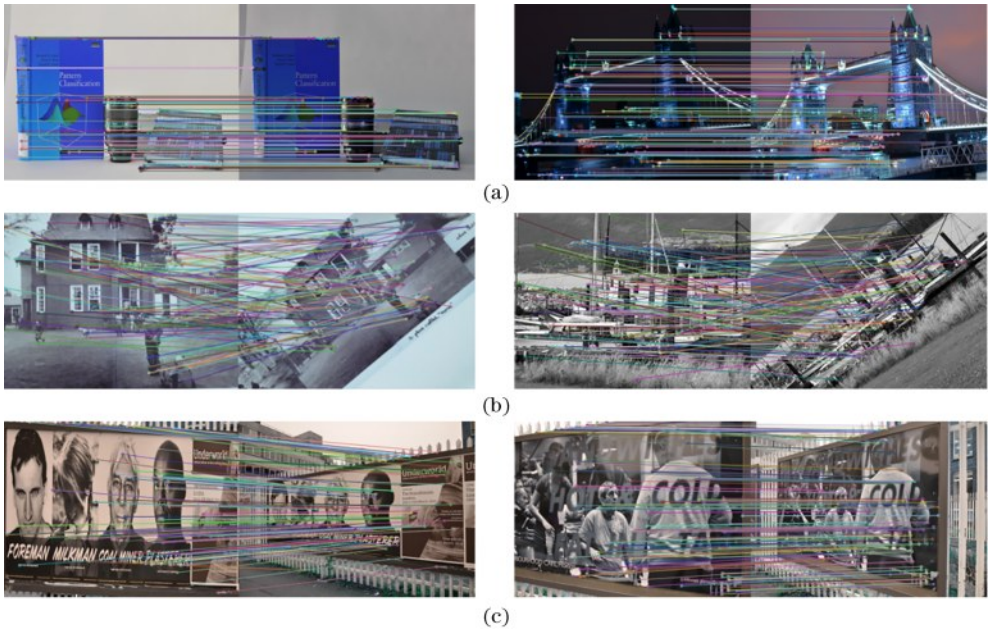

图 6　在 Hpatches 数据集上的可视化结果。(a)光照组；(b)旋转组；(c)视角组

Fig. 6　Qualitative results of proposed method on images pairs on HPatches dataset. (a)Illumination cases.; (b) rotation cases; (c) perspective cases

4　结　　论

提出了一种基于自监督学习的神经网络框架来训练关键点检测器和描述子检测器。与现有方法不同的是，主干网络采用残差网络结构提取不同尺度的特征，在描述子分支融合浅层局部特征和深层高级特征，得到分辨率更高的描述子特征，并引入通道注意力机制增强融合后的特征。在不显著增加计算量的前提下提升了描述子检测器性能，基于 HPatches 数据集的相关实验验证了所提方法的有效性。

参　考　文　献

[1] Cadena C, Carlone L, Carrillo H, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332.

[2] Agarwal S, Snavely N, Seitz S M, et al. Bundle adjustment in the large[M]//Daniilidis K, Maragos P, Paragios N. Computer vision-ECCV 2010. Lecture notes in computer science. Heidelberg: Springer, 2010, 6312: 29-42.

[3] Wang S, Clark R, Wen H K, et al. DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks[C]//2017 IEEE International Conference on Robotics and Automation, May 29-June 3, 2017, Singapore. New York: IEEE Press, 2017: 2043-2050.

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

[5] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF) [J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[6] Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF[C]//2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 2564-2571.

[7] Leutenegger S, Chli M, Siegwart R Y. BRISK: binary robust invariant scalable keypoints[C]//2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 2548-2555.

[8] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3852-3861.

[9] Yi K M, Trulls E, Lepetit V, et al. LIFT: learned invariant feature transform[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9910: 467-483.

[10] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]//NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, December 7-12, 2015, Montreal, Quebec, Canada. [S.l.: s.n.], 2015: 2017-2025.

[11] Ono Y, Trulls E, Fua P, et al. LF-Net: learning local features from images[C]//NIPS'18: Proceeding of the 32nd International Conference on Neural Information Processing System, December 3-8, 2018, Montreal, Quebec, Canada. [S.l.: s.n.], 2018: 6237-6247.

[12] 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 (CVPRW), June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 337-33712.

[13] Christiansen P H, Kragh M F, Brodskiy Y, et al. UnsuperPoint: end-to-end unsupervised interest point detector and descriptor[EB/OL]. (2019-07-09)[2021-07-06]. https://arxiv.org/abs/1907.04011.

[14] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[15] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[16] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[17] Tang J X, Kim H, Guizilini V, et al. Neural outlier rejection for self-supervised keypoint learning[EB/OL]. (2019-12-23) [2021-07-06]. https://arxiv.org/abs/1912.10615v1.

[18] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[19] Schmid C, Mohr R, Bauckhage C. Evaluation of interest point detectors[J]. International Journal of Computer Vision, 2000, 37(2): 151-172.

[20] Fischler M A, Bolles R C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography[J]. Communications of the ACM, 1981, 24(6): 381-395.