文章编号:1671-4598(2022)04-0222-07 DOI:10.16526/j.cnki.11—4762/tp.2022.04.037 中图分类号:TP391 文献标识码:A

# 结合多尺度与可变形卷积的自监督图像特征点提取网络

张少鹏[1]，周大可[1,2]，杨 欣[1,2]
(1. 南京航空航天大学 自动化学院，南京 210000；
2. 江苏省物联网与控制技术重点实验室，南京 210000)

**摘要**：特征点提取是图像处理领域的一个重要方向，在视觉导航、图像匹配、三维重建等领域具有广泛的应用价值；基于卷积神经网络的特征点提取方法是目前的主流方法，但由于传统卷积层的感受野大小不变、采样区域的几何结构固定，在尺度、视角和光照变化较大的情况下，特征点提取的精度和鲁棒性较差；为解决以上问题提出了一种结合多尺度与可变形卷积的自监督特征点提取网络；以 L2-NET 为网络骨干，在深层网络中引入多尺度卷积核，增强网络的多尺度特征提取能力，获得细粒度尺度信息的特征图；使用单应矩阵约束的可变形卷积以提取不规则的特征区域，同时降低运算量，并采用归一化约束单应矩阵的求解，均衡不同采样点对结果的影响，配合在网络中增加的卷积注意力机制和坐标注意力机制，提升网络的特征提取能力；文章在 HPatches 数据集上进行了对比试验和消融实验，与 R2D2 等 7 种主流方法进行对比，文章方法的特征点提取效果最好，相比于次优数据，特征点重复度指标（Rep）提升了约 1%，匹配分数（M. s.）提升了约 1.3%，平均匹配精度（MMA）提高了约 0.4%；文章提出的方法充分利用了可变形卷积提供的深层信息，融合了不同尺度的特征，使特征点提取结果更加准确和鲁棒。

**关键词**：特征点检测；多尺度卷积；可变形卷积；注意力机制

# A Self-supervised Feature Points Extraction Networks Based on Multi-scale and Deformable Convolution

ZHANG Shaopeng[1], ZHOU Dake[1,2], YANG Xin[1,2]
(1. College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China;
2. Jiangsu Key Laboratory of Internet of Things and Control Technologies, Nanjing 210000, China)

**Abstract**: Feature point extraction is an important direction in the field of image processing. It has wide application value in the fields of visual navigation, image matching, 3D reconstruction and so on. The feature point extraction method based on convolution neural network is the mainstream method at present. However, due to the constant size of the receptive field of the traditional convolution layer and the fixed geometric structure of the sampling area, the accuracy and robustness of feature point extraction are poor when the scale, viewing angle and illumination change greatly. A self-supervised feature points extraction network combining multi-scale and deformable convolution is proposed. Taking the L2-net as the backbone of the network, the multi-scale convolution kernel is introduced into the deep network to enhance the multi-scale feature extraction ability of the network and obtain the feature map of fine-grained scale information; The deformable convolution constrained by the homography matrix is used to extract irregular feature regions, reduces the amount of computation, and solves the normalized constrained homography matrix to balance the influence of different sampling points on the results, cooperates with the convolution attention mechanism and coordinate attention mechanism added in the network to improve the feature extraction ability of the network. In this paper, the comparative experiments and ablation experiments are carried out on the HPatches data set. Compared with seven mainstream methods such as R2D2, the feature point extraction effect of this method is the best. Compared with the suboptimal data, the feature point repeatability index (REP) is improved by about 1%, the matching score (M. s.) is improved by about 1.3%, and the accuracy of average matching (MMA) is improved by about 0.4%. The proposed method makes full use of the deep information provided by the deformable convolution, and integrates the features of different scales to make the feature point extraction results more accurate and robust.

**Keywords**: feature points extration; multi-scale convolution; deformable convolution; attention mechanism

**收稿日期**:2022-01-28； **修回日期**:2022-02-21。
**基金项目**:国家自然科学基金项目(61573182)。
**作者简介**:张少鹏(1997-),男,福建福州人,硕士研究生,主要从事计算机视觉方向的研究。
周大可(1973-),男,江苏淮安人,博士,副教授,主要从事计算机视觉方向的研究。
杨 欣(1978-),男,江苏镇江人,博士,副教授,主要从事计算机视觉方向的研究。
**引用格式**:张少鹏,周大可,杨 欣.结合多尺度与可变形卷积的自监督图像特征点提取网络[J].计算机测量与控制,2022,30(4):222-228.

## 0　引言

图像特征点检测任务是计算机视觉领域的底层任务，对许多更高层的任务，比如车辆检测[1]、视觉 SLAM[2]起到至关重要的支撑作用，更高的特征点检测精度能显著提升这些高层任务的性能。在早期研究中，特征检测任务主要通过手工设计特征提取策略和特征点描述策略来寻找特征点，比如 SIFT[3]、SURF[4]、AKAZE[5]、ORB[6]等。随着深度学习的兴起，卷积神经网络在计算机视觉领域开始广泛应用，如何利用深度学习进行特征点检测、描述和匹配也成为了研究的热点。根据深度学习在这些研究中起到的作用，可以将其分为三类：使用传统算法检测的特征点作为监督信息进行训练的方法，采用自监督特征描述网络＋手工筛选特征点的方法，和采用自监督方式同时训练特征点检测和描述网络的方法。

受采用人工提取策略的特征检测器启发，Verdie 等[7]将特征点检测问题视为回归问题，利用 SIFT 特征点来监督网络训练；Yi 等人[8]把特征点检测视为分类问题，预测每个像素点是特征点的概率，同样利用 SIFT 特征点监督训练，取得了不错的效果。但这类方法采用 SIFT 等传统策略提取的特征进行监督，导致特征点检测网络的上限被选用作为监督的人工设计检测器所限制，无法取得更优的效果。

卷积神经网络具有优异的图像特征提取能力，因此，一些研究者提出不训练特征检测网络，采用自监督的方式直接训练卷积神经网络提取输入图像每个像素局部区域的描述向量，然后人工设计筛选策略，选出潜在特征点[9-12]。Dusmanu 等人[13]等首先提出了这种方法，通过比较像素点和周围特征向量在通道域和空间域的差异来筛选特征。Tian 等人[14]等改进了筛选策略，以特征向量的方差大小和特征向量与周围像素的差异来挑选特征点。然而，这种方法在选取策略上还存在一定缺陷，虽然在光照变化下表现出了良好的性能，但在视角变化大时性能下降明显。

由于手工设计策略筛选特征点的方式仍存在对视角变化较敏感、难以达到全局最优等问题，许多研究选择训练卷积神经网络同时完成特征点的检测和描述。DeTone 等人[15]先训练网络提取三角形、四边形等简单图形中的角点，训练完成后再在真实图像上进行训练，使用输入图像多次进行单应矩阵扰动后提取的特征点作为伪真值标签。Ono 等人[16]将输入图像施加单应变换，以两张图像提取的特征点差异作为监督，同时加入空间变换网络，让网络提取的描述子具备空间不变性。Christina 等人[17]巧妙地将图像降采样与预测特征点位置相结合，同时输出特征点位置和描述向量。Barroso-Laguna 等人[18]采用人工设计的滤波器组预处理图像，再输入网络进行预测。Shen 等人[19]改进了 LF-Net 的网络结构以融合多尺度特征。在特征检测的过程中，重复纹理处提取的特征点常常会导致误匹配，所以需要针对性的处理重复纹理处提取的特征点。Revaud 等人[20]在网络中额外添加了评估描述子可信度的输出层，该层会在纹理重复度高的像素点位置输出一个较低的可信度，从而避免提取到这些特征点。然而，在特征点的提取过程中，卷积神经网络每一层的尺度变化较大，且只能以固定尺寸的卷积核对输入的特征图进行卷积，缺乏多尺度信息的融合；同时，图像局部的纹理形状并不规则，导致卷积层提取时可能引入无关信息、破坏局部纹理的完整性。所以如何融合多尺度的图像特征信息、更好地描述图像局部特征成为提高特征点提取网络性能的关键。

综上所述，本文针对以上问题，采用自监督的方式训练卷积神经网络，引入多尺度卷积，在同一个卷积层内融合多个不同感受野的卷积核，从而获得有丰富信息的多尺度特征，引入卷积注意力机制和坐标注意力机制与多尺度卷积核结合，使网络能够重点关注某一尺度特征图；引入可变形卷积，使网络输出的描述子可以灵活描述像素点周围的区域，最终获得了效果更佳的特征点检测网络。

## 1　方法

本文网络采取 L2-Net[21]作为基础网络，不同于 L2-Net 的输入为 32 * 32 的图像块，本文输入为整张图像，输出每个像素的描述向量、该像素是特征点的概率、该像素描述向量的可信度。在基础网络中加入多尺度卷积和卷积注意力模块，在网络最后额外增加两层单应约束的可变形卷积层和坐标注意力模块，网络整体结构如图 1 所示，输入图像经过共享特征融合模块后，从左到右依次输出特征点置信度响应图、特征点描述子、特征点局部区域重复度响应图。在训练时，本文预测两幅图像的特征点，并用光流预测的方法得到对应的特征匹配，并以此进行自监督训练。

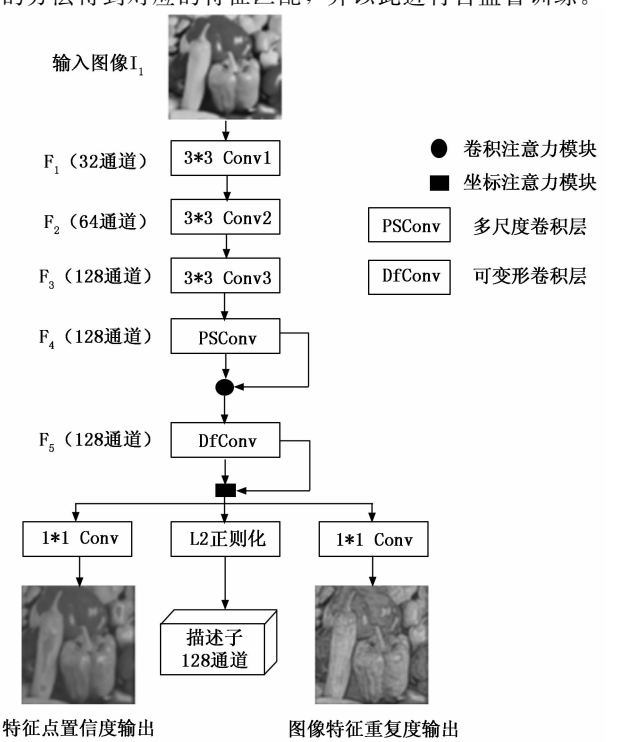


图 1　整体网络框架示意图

### 1.1 共享特征融合模块

本文首先通过一个共享的特征提取模块来提取后续三个输出模块需要的深度特征。为了获取足够丰富的特征信息，本文采用了多尺度卷积模块，该模块的输出通道与输入通道之间的关系如图 2 所示，不同颜色的方块代表不同空洞率的卷积，这部分输出特征通道由输入特征通道与该空洞率的卷积核卷积得到。首先通过四层 3 * 3 卷积层对图像进行预处理，再通过多尺度卷积层进行卷积，获得多尺度融合特征。但是，在网络中进行降采样会丢失信息，造成精度下降，因此，本文使用增加卷积空洞率的方式来代替降采样。

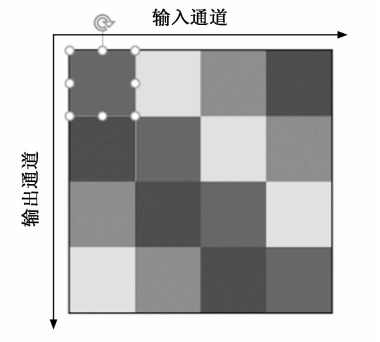


图 2 多尺度卷积层

PSConv[22]是 Li，Yao 等人提出的一种在同一层内提取多尺度特征的卷积层，其对输入的特征图采用不同空洞率的空洞卷积进行特征提取，空洞率沿着输入和输出通道的轴周期性地变化。卷积核的空洞率和卷积的感受野大小息息相关，空洞率越大，感受野越大。对于一个空洞率为 $d$，卷积核尺寸为 $K$ 的空洞卷积层，它的单层等效感受野 $R$ 的大小为：

$$R=(d-1)\times(K-1)+K \tag{1}$$

低空洞率的卷积更关注图像局部的信息，高空洞率的卷积能关注到更全局的图像信息。这种多尺度卷积设计使得网络能够在更细粒度范围内聚合不同尺度的特征图，有助于提取对尺度变化更鲁棒的特征点。

为提取到尺度不变的特征点，本文在多尺度卷积的基础上加入了卷积注意力机制（CBAM，convolutional block attention module）。CBAM 模块的结构如图 3 所示，通过依次应用通道和空间注意力来识别重要的特征区域。通过在多尺度卷积层后加入 CBAM 模块，网络能够在输入图像的场景尺度发生变化时提取到尺度相近的特征，有利于最后提取到尺度相对鲁棒的特征点。

### 1.2 归一化变换点对的单应约束可变形卷积层

可变形卷积是由 Zhu 等人[23]提出的一种卷积操作，设卷积核大小为 3，空洞率为 1 的传统卷积采样点为：

$$R=\{(-1,-1)\quad(-1,1)\quad\cdots\quad(1,-1)\quad(1,1)\} \tag{2}$$

$R$ 是以（0，0）为中心的周围 9 个点，则在每个卷积中心点 $p_0$ 输出的特征图 $y$（$p_0$）与输入特征图 $x$（$p_0$）的关系为：

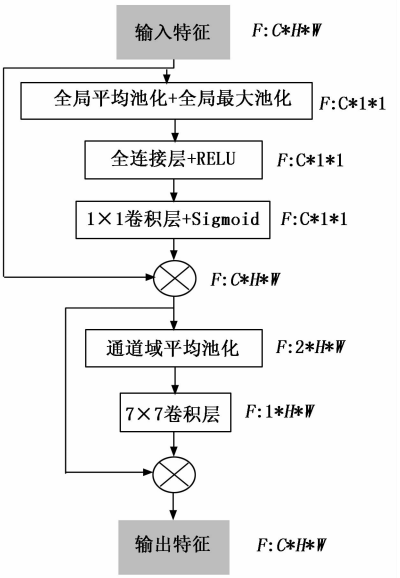


图 3 卷积注意力模块

$$y(p_0)=\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n) \tag{3}$$

其中：$w(p)$ 代表卷积核在点 $p$ 处的权重，$x(p)$ 代表输入特征图在点 $p$ 处的像素值。

而可变形卷积额外添加了两个卷积层用来预测采样点的偏移量 $\Delta p_n$ 和采样点的重要程度 $\Delta m_n$，其卷积公式为：

$$y(p_0)=\sum_{p_n\in R}w(p_n)\cdot x(p_0+p_n+\Delta p_n)\cdot\Delta m_n \tag{4}$$

通过预测偏移量，可变形卷积层可以采样到不规则的感兴趣区域轮廓，不必受限于规整的采样区域，能够增强网络输出的描述子的描述能力。

经典的可变形卷积核每个采样点都能在 $x$，$y$ 两个方向自由偏移，因此每个采样点的自由度为 2，一个卷积核的偏移量的自由度是 $2\cdot k^2$，$k$ 为卷积核大小。假设 $k=3$，则卷积核有 9 个采样点，输入图像大小为 $B\times C\times H\times W$，输出的预测偏移量大小应为 $B\times 18\times H\times W$。然而，在特征点检测任务中，完整的 $2\cdot k^2$ 自由度并不是必要的。在早期研究中，为了给网络提供空间不变性，研究者给网络添加了某些几何约束。Jaderberg 等人[24]通过空间变换网络（STN）预测一个二维仿射变换参数用于对齐输入图像，在 LF-Net[16]中，作者利用 STN 结构来变换输入图像，然后提取图像的描述子。这些研究直接估计输入图像的二维仿射变换，这种预测方法无法自适应调整局部的变换参数，同时仿射变换也无法完全描述二维射影图像间的几何关系。Luo[25]等人提出利用二维单应矩阵来约束可变形卷积的变化范围。定义一张图像上像素点的齐次坐标为 $(x,y,w)^T$，它属于射影空间 $IP^2$，射影映射是一种满足下列条件的可逆映射 $h$：$x_1,x_2,x_3$ 共线当且仅当 $h(x_1),h(x_2),h(x_3)$ 也共线[26]。射影映射组成的群称为射影变换群或单应变换群，存在一个非奇异矩阵 $H\in R^{3\times3}$ 满足：$h(x)=Hx$。

本文同样通过直接线性变换算法（DLT）来求解单应矩阵。设 $p_i=(x_i,y_i,w_i)^T$ 是卷积核的某个采样点，$p'_i=(x'_i,y'_i,w'_i)^T$ 是加上 offset 层预测的偏移量后的该点坐标，固定单应矩阵的平移为 0，最后一个元素 $h_{33}=1$，则有：

$$\begin{cases} A_ih=b_i \\ A_i=\begin{bmatrix} 0 & 0 & -w'_ix_i & -w'_iy_i & y'_ix_i & y'_iy_i \\ w'_ix_i & y'_ix_i & 0 & 0 & -x'_ix_i & -x'_iy_i \end{bmatrix} \\ h=[h_{11} \quad h_{12} \quad h_{21} \quad h_{22} \quad h_{31} \quad h_{32}]^T \\ b_i=[-y'_iw_i \quad x'_iw_i \quad \quad ]^T \end{cases} \tag{5}$$

每个坐标点可以提供两个约束，因此只需要预测 4 个坐标点的偏移量就可以求解出对应的单应矩阵，本文选取卷积核的 4 个角点进行预测。然而，由于 DLT 算法实质上是在最小化代数距离，当图像坐标发生变化时无法保证求解出的单应矩阵保持一致。进行归一化不仅可以提高结果的精度，还可以使 DLT 算法关于相似变换不变。因此在进行 DLT 求解前，本文还对相应坐标点进行了归一化，将选取的 4 个点对变换到标准坐标系，使 4 个采样点到坐标原点的距离为$\sqrt{2}$，求解后再对结果进行逆变换，得到原坐标系下的结果。

### 1.3　坐标注意力机制

通过观察特征点的分布，我们认为检测特征点不仅需要图像的局部信息，还需要网络能聚合更大范围内的图像空间信息。CBAM 注意力模块通过全局池化来引入位置信息，但这种方式只考虑了局部范围的信息。因此，本文采用坐标注意力机制[28]来捕获更大范围内的特征依赖关系。

如图 4 所示，从上层网络得到特征 $F\in R^{C\times H\times W}$，$C$，$H$，$W$ 分别是输入特征图的通道数、高和宽。对输入特征分别进行 $X$ 和 $Y$ 方向的一维全局平均池化操作得到，然后将其拼接在一起并依次通过一个 1×1 卷积核映射、一个 RELU 非线性激活层，再将两个方向的特征分离并分别通过一个 1×1 卷积核映射再与原输入特征做哈达玛积，这样每个注意力图都具备了沿着不同方向捕获输入特征图远距离依赖关系的能力。通过在编码器中加入坐标注意力机制可以获得可训练的空间权重，指导网络不局限于特征图的局部特征，提取到更丰富的特征，同时也能够加速训练收敛。

### 1.4　损失函数

通过自监督方式训练特征点检测网络实质上是通过合成新视图的方式获取真值。与 Revaud 等[20]提出的方法类似，本文的损失函数设置为：

$$L_{\text{total}}=\lambda_{\text{det}}L_{\text{det}}+\lambda_{\text{desc}}L_{\text{desc}} \tag{6}$$

其中：$L_{\text{det}}$用于训练特征点检测网络，$L_{\text{desc}}$用于训练描述子生成网络。$L_{\text{det}}$由余弦相似度损失 $L_{\text{cosim}}$和差异度损失 $L_{\text{peaky}}$构成。令 $S[p]$ 为网络预测的输入图像上点 $p$ 是特征点的可能性，$I$，$I'$ 分别是输入图像和合成的新视图，$P$ 是每个像素点周围区域，则余弦相似度损失为：

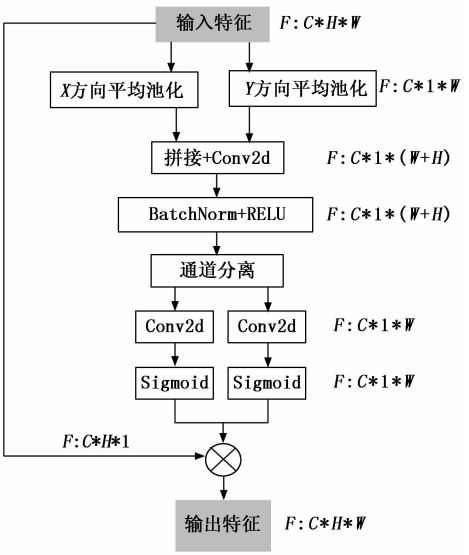


图 4　坐标注意力模块

$$L_{\text{cosim}}(I,I')=1-\frac{1}{|P|}\sum_{p\in P}\text{cosim}(S[p],S'[p]) \tag{7}$$

为保证选出的特征点在 $S$ 上的局部区域内有显著的大小差异，本文采用差异度损失：

$$L_{\text{peaky}}(I)=1-\frac{1}{|P|}\sum_{p\in P}\left(\max_{(i,j)\in p}S_{i,j}-\text{mean}_{(i,j)\in p}S_{i,j}\right) \tag{8}$$

综上，检测子损失为：

$$L_{\text{det}}=L_{\text{cosim}}(I,I')+0.5(L_{\text{peaky}}(I)+L_{\text{peaky}}(I')) \tag{9}$$

描述子损失 $L_{\text{desc}}$用于衡量两幅图像中同一个特征点生成的描述子是否相似，使用可微分的平均精度损失（AP）[29]：

$$\text{Prec}K=\frac{1}{K}\sum_{i=1}^{K}1[x_i\in S_q^+]$$

$$AP(q)=\frac{1}{|S_q^+|\sum_{K=1}^{n}1[x_K\in S_q^+]\text{Prec}K}$$

$$L_{AP}(q)=1-AP(q) \tag{10}$$

其中：$x\in\{x_1,x_2,\cdots,x_n\}$是输入图像 $I$ 中某个特征点与合成图像 $I'$ 中所有特征点进行描述子距离计算的结果，从 1 到 $n$ 描述子距离依次增加，根据它们是否是特征点的正确匹配，$x$ 可分为 $S_q^+$，$S_q^-$ 两个子集，则有 $1[x_i\in S_q^+]=\begin{cases}1 & x\in S_q^+\\ 0 & x\in S_q^-\end{cases}$，Prec$K$ 表示 $x$ 序列前 $K$ 个匹配中正确匹配的占比，$AP(q)$ 的物理意义为，当第 $K$ 个匹配是正确匹配时，比它与输入图像的特征点描述子距离更近的匹配中是正确匹配的占比大小。通过优化 $L_{AP}(q)$，可以监督正确匹配的特征点之间的描述子差异减小。

与网络预测的描述子可信度结合，最终的描述子损失函数为：

$$L_{AP_\kappa}(i,j)=1-[AP(i,j)R_{ij}+\kappa(1-R_{ij})] \tag{11}$$

其中：$R$ 是网络输出的各像素点描述子可信度，$\kappa$ 是超参数。

## 2 实验结果与分析

### 2.1 实验设置

2.1.1 数据集

使用图像检索数据集 Achen[30] 和 Random Web Image 数据集作为训练数据，Hpatches[31] 数据集作为测试数据集，以便与现有工作进行比较。Aachen 数据集包含 6 697 张参考图像，1 015 张明暗不同的查询图像以及对应的位置信息。原始图像大小为 1 600×1 063 像素，为降低网络计算量，将输入图像的大小设置为 640×480 像素。在选定图像对后，通过 SFM 和稠密光流预测，可以获得对应的位姿变换和两张图像上逐像素对应坐标，在训练中，网络预测获得图像对的特征点后，通过光流可以获得特征点在另一张图像上的位置，可以据此进行自监督训练。

2.1.2 训练细节

使用的训练平台为英伟达 RTX 2080，整体算法采用 pytorch 框架实现。输入图像大小为 640×480 像素，采用随机水平翻转、亮度调整、对比度调整、小幅度射影变换扰动来实现数据增强，采用 Adam 作为优化算法，学习率设置为 0.001，batch 大小为 4，共训练 20 个 epoch，权值衰减为 0.000 5。

2.1.3 评估指标

本文使用评估特征点检测精度最常用的几个指标，包括匹配分数（MS，matching score），即正确匹配在两图像共视特征点中的占比，它能反映提取的特征点质量；

$$M.s = \frac{1}{N}\sum_{i=1}^{N}\frac{M_{\text{inlier}}^{i}}{Num_{\text{co}}^{i}} \tag{12}$$

其中：$N$ 是测试集样本数量，$M_{\text{inlier}}$，$Num_{\text{co}}$ 分别是正确匹配数量和图像中共视特征点数量。

平均匹配精度（MMA，mean matching accuracy），即正确匹配特征点与所有可能匹配之比，MMA 越高，说明特征点的匹配正确率越高，反映了特征点描述子的区分能力。

$$MMA = \frac{1}{N}\sum_{i=1}^{N}\frac{M_{\text{inlier}}^{i}}{M_{\text{total}}^{i}} \tag{13}$$

其中：$M_{\text{total}}$ 是两幅图像的特征点采用暴力匹配得到的所有可能匹配数量。

特征点重复度（Rep，keypoint repeatability），即所有可能匹配与两图共视特征点数量之比，它能反映网络对某个纹理区域的特征提取能力。

$$Rep = \frac{1}{N}\sum_{i=1}^{N}\frac{M_{\text{total}}^{i}}{Num_{\text{co}}^{i}} \tag{14}$$

单应矩阵估计精度（HA，homography accuracy）：使用检测特征点估计的单应矩阵与两图单应矩阵真值之间的差距，由于特征点匹配常用于三维重建、导航等需要求解变换矩阵的场景，因此该指标也很重要。

$$HA = \frac{1}{N}\sum_{i=1}^{N}\frac{H_{\text{pred}}}{H_{\text{gt}}} \tag{15}$$

### 2.2 实验结果

2.2.1 对比实验

我们在 HPatches 数据集上比较了本文方法和 7 种主要方法，如 R2D2[20]、D2Net[13]，SIFT[3] 等，各对比方法的结果数据来自原论文。表 1 给出了在 3 像素误差阈值下 8 种方法的实验结果，其中，各对比方法的结果数据来自原论文。由于部分用于对比的方法在论文中缺乏相关指标的实验结果数据，且代码未开源，因此，表 1 中有部分数据缺失。结果表明，本文方法在大部分指标上的评估结果优于其他 7 种方法。相较于其他方法，本文采用的方法在重复度指标上较次优方法提升了约 1%，对于同一个纹理区域，本文方法能提取出的有效特征点更多，特征提取能力更佳；匹配分数上较次优方法提升了约 1.3%，说明采用本文方法提取的特征点提取质量更好；平均匹配精度提高了约 0.4%，说明采用本文方法提取的特征点在匹配时误匹配更少，网络提取的描述子能更好地区分不同特征点。虽然单应矩阵估计精度未取得最优效果，但仍能达到较优秀的精度。

表 1 在 HPatches 数据集上结果比较

| 方法 | Rep | M. s | MMA | HA |
|---|---|---|---|---|
| ORB（Rublee 等 2011） | 0.525 | 0.204 | / | 0.607 |
| AKAZE（Alcantarilla 等 2011） | 0.572 | 0.289 | / | 0.743 |
| SIFT（Lowe 等 2004） | 0.421 | 0.265 | / | 0.71 |
| LF-NET（Christiansen 等 2019） | 0.523 | 0.241 | / | 0.67 |
| superpoint（DeTone 等 2018） | 0.581 | 0.281 | 0.65 | **0.864** |
| D2-Net（Dusmanu 等 2019） | 0.317 7 | 0.204 | 0.4 | 0.406 |
| R2D2（Revaud 等 2020） | 0.643 | 0.292 | 0.721 | 0.670 |
| 本文 | **0.649** | **0.296** | **0.724** | 0.726 |

为了更好地展示本文方法在不同情况下的性能，图 5 展示了不同方法在光照、视角变化情况下 MMA 指标随像素误差阈值变化的图像，最左侧图像为两种变化下 MMA 指标的平均值。由于图 5 中部分方法仅有 HPatches 数据集下 MMA 这一指标的实验数据，因此这部分方法未在表 1 中列出。如图 5 所示，当像素误差阈值上升到 2px 时，本文方法的 MMA 指标综合上看已取得了最优效果，随着阈值限制的放宽，在不同特定条件下的效果也取得了最优效果。综合不同曲线，本文方法在大多数情况下取得了最优的匹配精度。

图 6 是在 HPatches 数据集中分别选取光照、视角变化较大的两个场景进行特征点匹配测试，上行为光照变化，下行为视角变化，像素误差阈值为 3px，每张图像提取 4 000个特征点，提取图像对的特征点后，通过数据集提供的图像对之间的变换矩阵将一幅图像中的特征点投影到另一幅图像中，并取投影位置周围 3 个像素范围内的特征点进行特征匹配，最后输出匹配结果。从图 6 可以看出，在光照条件变化大的情况下，本文方法提取的特征比传统的 SIFT、ORB 等特征有更多的正确匹配特征点对，说明匹配效果更好，相比 D2-Net、R2D2 等采用深度学习的方法，本文的方法更少在图像的低信息区域提取特征；在视角变化

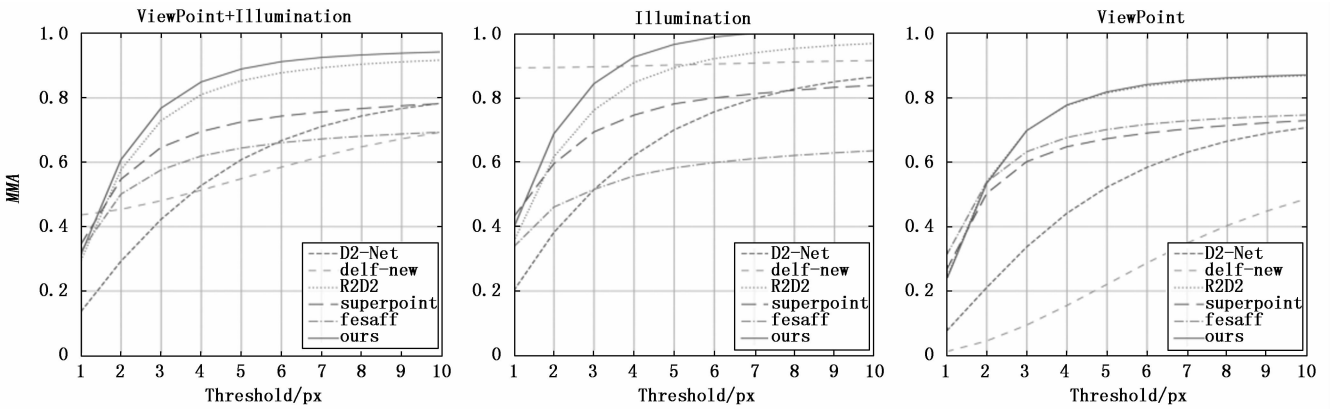


图 5　HPatches 数据集下部分方法平均匹配精度（*MMA*）随像素误差阈值变化曲线

时，相较于其他方法，本文方法提取的特征分布更加均匀，定位也更精确。在曲线的前半段可以看出，仅需稍微放宽像素误差阈值的限制，本文方法获得的精度提升最大，更容易达到最优效果。

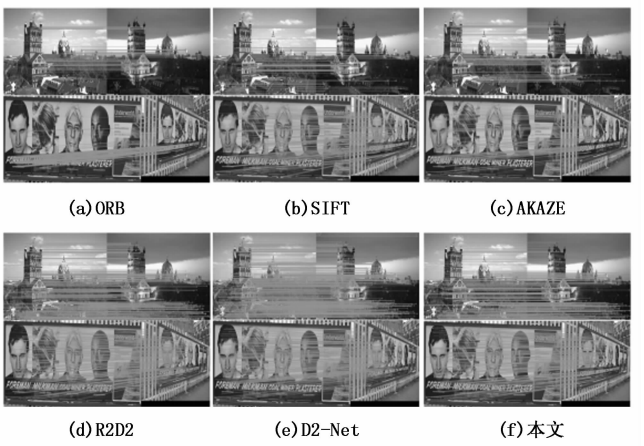


图 6　不同方法在光照条件、视角变化下进行特征点提取、匹配的结果

### 2.2.2　消融试验

本文采用 L2-Net 作为基准网络，在此基础上加入了多尺度卷积层、可变形卷积层和卷积注意力、坐标注意力模块，为进一步探究各模块对实验结果的影响，分别去除注意力机制、CBAM＋多尺度卷积模块、CA＋可变形卷积模块，将得到的评估指标进行比较，结果如表 2 所示。实验结果表明，增加注意力机制和不同的卷积模块均会带来一定的精度提升，可变形卷积模块能有效提升。

表 2　HPatches 数据集上消融实验结果比较

| 方法 | Rep | M. s | MMA | HA |
|---|---|---|---|---|
| 去除注意力机制 | 0.639 | 0.291 | 0.718 | 0.713 |
| CBAM＋多尺度卷积 | 0.641 | 0.290 | 0.722 | 0.670 |
| CA＋可变形卷积 | 0.644 | 0.293 | 0.720 | 0.721 |
| 本文 | 0.649 | 0.296 | 0.724 | 0.726 |

## 3　结束语

为改进特征点检测及描述算法的性能，本文提出了一种端到端的自监督特征点提取网络。在 L2-NET 网络基础上，通过加入多尺度卷积和可变形卷积，优化可变形卷积的单应约束求解步骤，并引入卷积模块注意力和坐标注意力机制，提升网络对特征点的检测能力。实验结果表明，本文方法实现了更高的特征点检测精度，同时对光照、视角变化有更强的鲁棒性，在各类场景中均有较高的应用价值。

**参考文献：**

[1] 李忠海，李建伟. 基于特征点光流聚类的复杂背景中运动车辆检测 [J]. 计算机测量与控制，2016，24 (5)：234－236.

[2] 牛文雨，李文锋. 基于动态物体特征点去除的视觉里程计算法 [J]. 计算机测量与控制，2019，27 (10)：218－222.

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

[4] BAY H，TUYTELAARS T，GOOL L V. SURF：Speeded up robust features [C] // Proceedings of the 9th European Conference on Computer Vision-Volume Part I. Springer-Verlag，2006.

[5] ALCANTARILLA P F. Fast explicit diffusion for accelerated features in nonlinear scale spaces [C] // British Machine Vision Conference (BMVC)，2013.

[6] RUBLEE E，RABAUD V，KONOLIGE K，et al. ORB：an efficient alternative to SIFT or SURF [C] // IEEE International Conference on Computer Vision，ICCV 2011，Barcelona，Spain，IEEE，2011.

[7] VERDIE Y，KWANG M Y，FUA P，et al. TILDE：a temporally invariant learned detector [C] //2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，Boston，MA，USA：IEEE，2015：5279－5288.

[8] YI K M，TRULLS E，LEPETIT V，et al. LIFT：Learned Invariant Feature Transform [M]. Springer International Pub-

lishing, 2016.
[9] TIAN Y, YU X, FAN B, et al. SOSNet: second order similarity regularization for local descriptor learning [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019.
[10] LUO Z, SHEN T, ZHOU L, et al. ContextDesc: local descriptor augmentation with cross-modality context [J]. IEEE, 2019.
[11] SIMO-SERRA E, TRULLS E, FERRAZ L, et al. Discriminative learning of deep convolutional feature point descriptors [C] // IEEE International Conference on Computer Vision, IEEE, 2016.
[12] MA T, WANG Y, WANG Z, et al. ASD-SLAM: a novel adaptive-scale descriptor learning for visual SLAM [C] // 2020 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2020.
[13] DUSMANU M, ROCCO I, PAJDLA T, et al. D2-Net: A trainable CNN for joint description and detection of local features [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019.
[14] TIAN Y, BALNTAS V, NG T, et al. D2D: keypoint extraction with describe to detect approach [M]. Cham: Springer International Publishing, 2021.
[15] 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), Salt Lake City, UT, USA: IEEE, 2018: 33710 - 33712.
[16] ONO Y, TRULLS E, FUA P, et al. LF-Net: learning local features from images [J/OL]. ArXiv: 1805.09662 [Cs], 2018 [2021-03-10]. http://arxiv.org/abs/1805.09662.
[17] CHRISTIANSEN P H, KRAGH M F, BRODSKIY Y, et al. UnsuperPoint: end-to-end unsupervised interest point detector and descriptor [J/OL]. ArXiv: 1907.04011 [Cs], 2019 [2021-03-13]. http://arxiv.org/abs/1907.04011.
[18] BARROSO-LAGUNA A, VERDIE Y, BUSAM B, et al. HDD-Net: hybrid detector descriptor with mutual interactive learning [C] //Proceedings of the Asian Conference on Computer Vision, 2021.
[19] SHEN X, WANG C, LI X, et al. RF-Net: an end-to-end image matching network based on receptive field [J/OL]. ArXiv: 1906.00604 [Cs], 2019 [2021-03-18]. http://arxiv.org/abs/1906.00604.
[20] REVAUD J, WEINZAEPFEL P, SOUZA C R de, et al. R2D2: repeatable and reliable detector and descriptor [C] // NeurIPS, 2019.
[21] TIAN Y, FAN B, WU F. L2-Net: deep learning of discriminative patch descriptor in euclidean space [C] //2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, 2017: 6128 - 6136.
[22] LI D, YAO A, CHEN Q. PSConv: squeezing feature pyramid into one compact poly-scale convolutional layer [C] //Computer Vision - ECCV 2020, Cham: Springer International Publishing, 2020: 615 - 632.
[23] ZHU X, HU H, LIN S, et al. Deformable ConvNets V2: more deformable, better results [C] //2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019.
[24] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. spatial transformer networks [C] //Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Cambridge, MA, USA: MIT Press, 2015: 2017 - 2025.
[25] LUO Z, ZHOU L, BAI X, et al. Aslfeat: Learning local features of accurate shape and localization [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6589 - 6598.
[26] HARTLEY R, ZISSERMAN A. Multiple view geometry in computer vision [M]. Cambridge university press, 2003.
[27] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713 - 13722.
[28] HE K, LU Y, SCLAROFF S. local descriptors optimized for average precision [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018.
[29] SATTLER T, WEYAND T, LEIBE B, et al. Image retrieval for image-based localization revisited [C] // British Machine Vision Conference, 2012.
[30] BALNTAS V, LENC K, VEDALDI A, et al. HPatches: a benchmark and evaluation of handcrafted and learned loca [C] //IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017: 3852 - 3861.

(上接第 221 页)

[13] MENG Y, ZHA J, LIU Y. Intensifying the SNR of BOTDA using adaptive constrained least squares filtering [J]. Optics Communications, 2019, 437 (18): 219 - 225.
[14] ZHAO Y, SUN X, ZHANG C, et al. Using Markov constraint and constrained least square filter to develop a novel method of passive terahertz image restoration [J]. Journal of Physics: Conference Series, 2019, 1187 (4): 2282 - 2292.
[15] DAN C, HUIFANG X, JINWU X. An improved Richardson-Lucy iterative algorithm for C-scan image restoration and inclusion size measurement [J]. Ultrasonics, 2018, 91 (13): 103 - 113.
[16] FAN W S, WANG H Y, WANG Y, et al. Blind deconvolution with scale ambiguity [J]. Applied Sciences-Basel, 2020, 10 (3): 939 - 957.
[17] HUANG C, CHEN F, CHANG Y, et al. Adaptive operator-based spectral deconvolution with the Levenberg-Marquardt algorithm [J]. Photonic Sensors, 2020, 10 (3): 242 - 253.