引用格式：赵艳明，林美秀，曾姝瑶．AttentionRanker－基于排名优化的自－互注意力机制[J]．中国传媒大学学报（自然科学版），2023，30（04）：27-38.
文章编号：1673-4793（2023）04-0027-12

AttentionRanker——基于排名优化的自－互注意力机制

赵艳明，林美秀*，曾姝瑶*
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：图像匹配是精准估计相机位姿信息的关键，近年来基于深度学习注意力机制的图像匹配研究取得了较大进展，但如何降低Transformer类图像匹配网络的高计算复杂度仍是巨大挑战。为了提高匹配网络效率，本文提出一种基于排名优化的自－互注意力机制。通过对位置编码后的一维输入特征图重塑形，采用类空间注意力机制挑选Top－m个活跃像素点的方法稀疏注意力图，成功地将点积注意力的时间复杂度从二次降为近线性。实验结果表明该方法在前向推理时耗时更短，并且能在一定程度上提升位姿估计精度。

关键词：图像匹配；注意力机制；稀疏算法
中图分类号：TP183　**文献标识码**：A

AttentionRanker——self－cross attention mechanism based on ranking optimization

ZHAO Yanming, LIN Meixiu*, ZENG Shuyao*
（School of Information and communication Engineering, Communication University of China, Beijing 100024,China）

Abstract: Image matching is the key to accurate camera pose estimation. In recent years, the research on image matching based on the attention mechanism of deep learning has made great progress, but it is still a great challenge to reduce the high computational complexity of Transformer-like image matching networks. In order to improve the matching network efficiency, in this paper a self-cross attention mechanism based on ranking optimization was proposed. By reshaping the one-dimensional input feature map after position encoding and using a spatial-like attention mechanism to pick Top-m active pixel points to sparse the attention map, the time complexity of dot product attention was successfully reduced from quadratic to nearly linear. Experimental results show that the method is less time consuming in forward inference and can improve the accuracy of pose estimation to a certain extent.

Keywords: image matching; attention mechanism; sparse algorithm

1　引言

图像匹配在40年前由David Marr[1]教授首次提出，旨在探索不同视觉对象之间的差异性和共同性，并且作为计算机视觉的底层任务连接着两个具有相同或相似属性的图像目标，是计算机视觉中最为重要的研究领域之一。

相机位姿估计任务作为图像匹配的一个基础下游

基金项目：广播电视和网络视听中长期科技计划项目（2022AF0300）
作者简介（*为通讯作者）：赵艳明（1973-），女，博士，副教授，主要从事计算机三维视觉研究。email: yanmingzhao@cuc.edu.cn；林美秀（2003-），女，本科生，主要从事计算机三维视觉研究。Email：2581652378@qq.com；曾姝瑶（1998-），女，硕士研究生，主要从事计算机三维视觉研究。Email：zsyao65@outlook.com

任务,需要匹配网络提供对应的点对匹配信息从而还原出相机的旋转平移运动,如图1所示,它作为低层视觉通往高层视觉的纽带,不但承接着三维重建、同步定位与地图构建(Simultaneous Localization and Mapping, SLAM)等大型任务,同时也是实现信息识别与整合[2-4]以及从低维图像恢复高维结构[5-6]的重要途径。

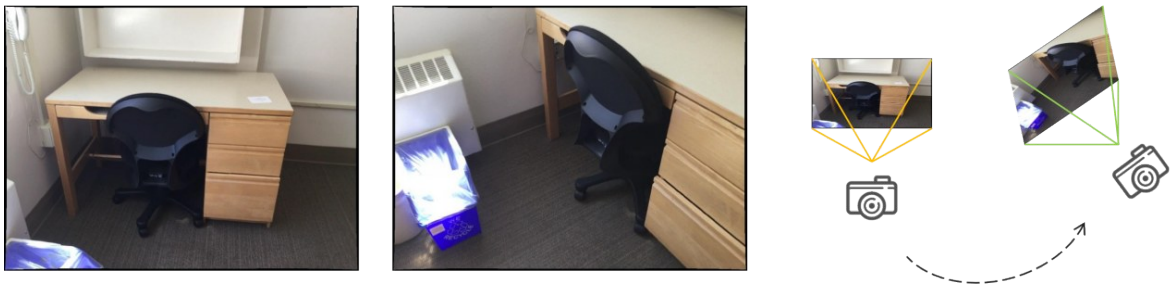


图1 位姿估计任务示意图

目前大多数图像匹配算法通常包括三个独立的步骤:特征检测、特征描述和特征匹配。近年来随着深度学习的迅速发展,这三个步骤逐渐被整合到一个端到端网络当中,利用深度神经网络根据不同图像集特点在特征检测阶段学习到特征点之间的关系并进行匹配。然而由于很多室内数据集图像中的弱纹理区域或者重复区域往往会占据图像的大部分空间,并且相机运动和光照条件会带来强视点变化和强光线变化,这使得特征检测器很难提取到可重复的特征点,从而无法找到正确的特征点对应关系。最近的一些研究工作直接通过建立像素级的密集匹配并在其中选择置信度高的匹配对,避免了特征检测器无法提取到足够多的特征点进行后续匹配的问题。

针对原始Transformer结构处理长序列时带来的显存爆炸问题,虽然已经有很多研究提出了高效的Transformer变体,但其中绝大多数研究集中于自然语言处理的稀疏方法,在计算机视觉领域则通常直接引用前者思路,缺少针对性面向图像处理的注意力稀疏算法。

围绕上述问题,本文展开研究工作,通过梳理自-互注意力机制在提取得到的密集局部特征中进行信息交互的过程,提出了基于排名优化的自-互注意力方法-AttentionRanker。该算法创新性地通过对位置编码后的一维输入特征图进行重塑形,然后利用类空间注意力机制挑选少量活跃像素点,成功地将每层注意力的时间复杂度降为$O(N\cdot \ln N)$,对于不同图像生成不同的权值从而实现自适应优化。

2 相关工作

2.1 无特征检测器的图像匹配算法研究现状

密集特征匹配思想可以追溯到2010年的Liu等人[7]提出的基于光流法的SIFT Flow。2018年Ignacio等人[8]针对弱纹理区域和图案重复区域用最近邻方法容易产生错误匹配的问题,提出邻域共识网络(Neighbourhood Consensus Network, NC-Net),它通过构造4D代价容量函数来枚举图像之间所有可能的匹配点对,然后利用4D卷积对代价容量进行正则化,以邻域共识思想约束所有的匹配点对。然而NC-Net中的4D卷积神经网络也带来了巨大的内存消耗和时间复杂度问题,2020年,Li等人[9]提出的双分辨率对应网络(Dual-Resolution Correspondence Networks, DRC-Net)同样通过构造四维卷积神经网络获取密集匹配,通过这种由粗到细的方式极大地提高了匹配的可靠性并且避免了整个网络都进行4D卷积运算所带来的巨大计算代价。

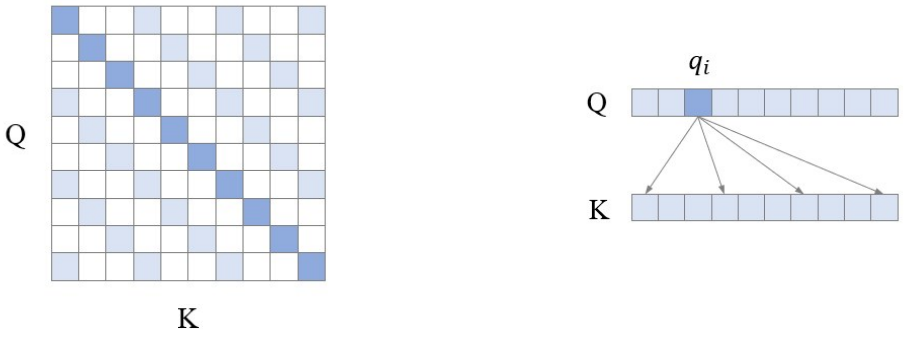
2021年CVPR挑战赛中Sun等人[10]提出了在SuperGlue[11]的匹配思路下设计的基于Transformer的图像匹配网络LoFTR[10]。其整体可分为四个组成部分:特征金字塔、自-互注意力信息传递、粗匹配预测、多尺度特征融合匹配。

首先输入两张图片I^A、$I^B\in\mathbb{R}^{h\times w}$,然后构建一个具有三层结构的ResNet-FPN网络,输出粗精度特征图$\tilde{F}$和细精度特征图$\hat{F}$。然后将得到的一对粗精度特征图分别展平为一维向量$\tilde{F}^A$、$\tilde{F}^B\in\mathbb{R}^{N\times d}$,融合位置编码后送入自-互注意力模块,得到图像内部的关键点信息以及图像之间的关键点信息。然后利用Sinkhorn算法[12-13]或双Softmax(Dual-softmax)法得到粗精度匹配预测。最后是进行多尺度特征融合匹配,对于每一对粗匹配(i,j),在细精度特征图$\hat{F}$上定位其位置,然后裁剪两组大小为$w\times w$的网格窗口并展平,通过自-互注意力信息传递后,得到两个以粗匹配预测的定位点i和j分别作为$\hat{F}^A$和$\hat{F}^B$中心的细精度局

部特征表示。通过计算概率分布的期望，收集 $\hat{F}^A$ 中所有特征点的对应匹配后，最终得到细精度特征图上的亚像素级匹配 $(i,j') \in M_f$。

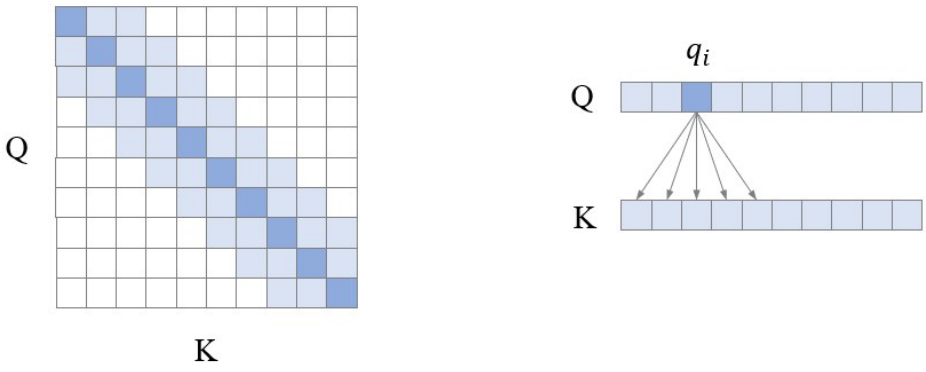
2.2 注意力矩阵的稀疏分解

为了降低注意力模型的时间复杂度，Zaheer 等人[14]提出了两个假设的注意力模型，分别是空洞注意力模型（图2）和局部注意力模型（图3），这两种模型在计算上都有所简化。



(a) 空洞注意力矩阵　　(b) 空洞注意力元素间的关联关系

图2 空洞注意力的注意力矩阵及其关联关系示意图



(a) 局部注意力矩阵　　(b) 局部注意力元素间的关联关系

图3 局部注意力的注意力矩阵及其关联关系示意图

与这两种算法有相似之处，Sparse Transformer[15]在注意力的计算上直接将两个假设合并起来，也就是对于每一个元素来说，都只和与它距离不超过 k，以及距离为 $mk(k>1)$ 的元素相关联，这样不仅可以学习紧密相关的局部信息，并且在全局关联性的计算中稀疏了一些注意力，降低计算复杂度。具体算法如下：

定义一个集合 $S=S_1,\cdots S_N$，N 为向量长度。S_i 表示第 i 个输出向量对应于输入向量中的索引集合，即第 i 个元素可以关联到的元素集合，输入向量 X 通过 S 与输出向量关联起来（公式（1）、（2））：

$$corr(X,S)=\left(attention(x_i,S_i)\right)_{i\in\{1,\cdots,N\}} \tag{1}$$

$$attention(x_i,S_i)=softmax\left(\frac{(W_q x_i)K_{S_i}^T}{\sqrt{d}}\right)V_{S_i} \tag{2}$$

其中 $K_{S_i}=W_k x_j, V_{S_i}=W_v x_j (j\in S_i)$，$W_q$、$W_k$、$W_v$ 分别表示将给定输入元素 x_i 转换为 query、key 和 value 的权重矩阵，$attention(x_i,S_i)$ 表示 x_i 和可以关注的元素之间的注意力。

当使用两个注意力头时，让每个注意力关注不同的位置，文中选取让其中一个注意力头只关注当前位置的距离为 $k=\sqrt{N}$ 以内的元素，让另一个注意力头只关注距离当前位置为 $mk=m\sqrt{N}$ 的元素。这样就将计算复杂度由 $O(N^2\cdot d)$ 降低为 $O\left(N\sqrt{N}\cdot d\right)$。

3 本文方法

3.1 自互注意力机制

图像匹配任务的传统方法是在获取特征点后计算其视觉描述符信息,然后通过暴力匹配计算描述符欧氏距离获得匹配点对。近年来的匹配算法受Transformer[16]的启发,在图神经网络的基础上,利用注意力机制整合其他的上下文线索,从而给特征点或者特征图赋予更多的全局信息。

使用卷积神经网络提取两张原始图像I^A、$I^B \in \mathbb{R}^{h\times w}$的局部特征图$F^A$和$F^B$,自-互注意力模块提取密集匹配过程如下:

(1)使用绝对正弦-余弦位置编码为F^A和F^B中的每个元素添加特定的位置信息,使得图像上的特征与其所在的位置相关联,提高在弱纹理区域找到对应匹配区域的能力。参考Carion等人[17]的位置编码方法,将第i个特征通道中(x,y)位置的正弦-余弦位置编码的二维扩展$PE^i_{x,y}$定义为式(3):

$$PE^i_{x,y}=\begin{cases}\sin(\omega_k\cdot x), & i=4k\\ \cos(\omega_k\cdot x), & i=4k+1\\ \sin(\omega_k\cdot y), & i=4k+2\\ \cos(\omega_k\cdot y), & i=4k+3\end{cases},\ k\in\left[0,\frac{h\times w}{4}\right] \tag{3}$$

其中$\omega_k=1\Big/10000^{\frac{2k}{d}}$,$d$是使用了位置编码后的特征通道数。

(2)将特征图F^A和F^B展平为一维向量,分别与位置编码融合相加得到F^A_{pe}和F^B_{pe}后输入自-互注意力模块。

(3)对两个序列计算图注意力:对于自注意力层,输入特征f_i和f_j相同,来自于F^A_{pe}或F^B_{pe};对于互注意力层,输入特征f_i和f_j则分别来自于F^A_{pe}和F^B_{pe}(或者F^B_{pe}和F^A_{pe},具体情况视互注意力方向而定)。

(4)将自-互注意力模块中的自注意力层和互注意力层交替N_c次,对输入特征进行变换,最终输出融合本张图片邻域信息与待匹配图像信息的特征F^A_{tr}和F^B_{tr}。

图4给出了基于Transformer的自-互注意力流程。自注意力层使得每个点关注其周围所有点以及关联性,互注意力层使得每个点关注另一幅图上的所有点及其关联性。

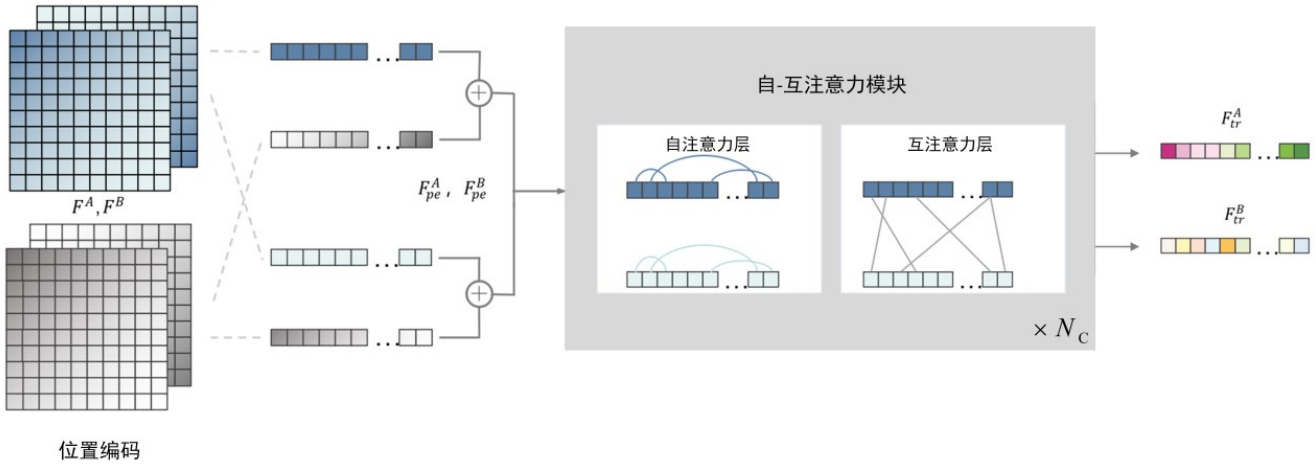


图4 特征图F^A、F^B的自-互注意力流程

3.2 基于排名优化的自-互注意力机制

因为直接使用普通的Transformer编码器结构对算力要求过高,为了能够轻量化使用Transformer,本小节根据输入图像的不同特点进行针对性处理,结合活跃像素点的注意力挑选策略,提出基于排名优化的自-互注意力机制。

3.2.1 活跃像素点的挑选策略

针对普通注意力机制中忽略稀疏性,对所有的query和key进行点积运算从而造成时间复杂度高的问题,一方面需要考虑不遗漏计算重要的注意力,另一方面需要考虑如何有效地减少计算量。对于每一个一维向量$F_{pe}\in\mathbb{R}^{N\times d}$,通过线性映射后得到查询向量$q\in\mathbb{R}^d$、值向量$k\in\mathbb{R}^d$和键向量$v\in\mathbb{R}^d$。如图5所示,本节跟随Informer[18]的实验结论定义两种查询类型,活跃查询q_a(active query)和非活跃查询q_l(lazy query):

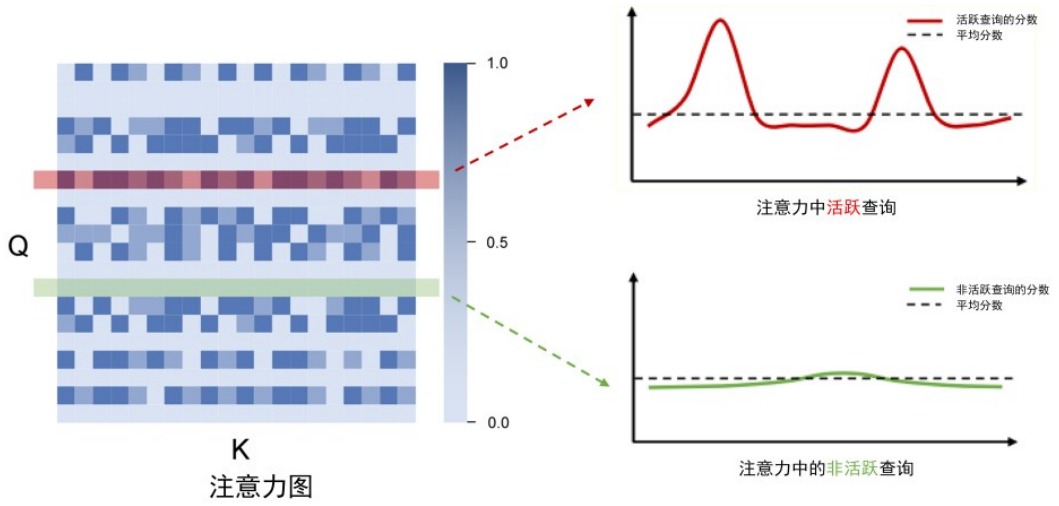


图5 活跃查询和非活跃查询的注意力分布示意图

（1）q_a是能在key中查询出更关键的信息的query，即q_a-key点积对对于注意力有贡献，这种query在注意力中有一个或多个注意力分数的峰值，其他地方的分数则比较低。

（2）q_l是使key起平均值作用的query，即q_l-key点积对对于注意力仅仅起很微弱的贡献。这种query在注意力中注意力分数没有太大的起伏，整体分布比较平均。

为了从所有query中量化区分“活跃性”，在每次进入自注意力层和互注意力层之前首先将一维向量进行重新整合，转换为特征图大小的向量$x \in \mathbb{R}^{(h \times w) \times d}$，此时的隐藏维度$d$可以看作是通道数，图像上的每个像素点经过特征提取和位置编码融合后，使得x不但带有丰富的位置信息，且携带了特征的抽象表达，而这种抽象表达的信息更多体现在“通道维度”上。

如图6所示，利用Woo等人[19-20]提出的空间注意力思想，对通道进行降维操作，将隐藏维度带有的信息压缩后送入类空间注意力模块，实现对特征图x的重构。特征图x同时经过全局平均池化[21]和全局最大池化，得到两种不同的通道特征描述算子后将其进行拼接：

$$x' = Concat\left[Avgpool(x),\ Maxpool(x)\right] \quad (4)$$

其中特征图$x' \in \mathbb{R}^{(h \times w) \times 2}$。将拼接得到特征图$x'$经过输出通道数$out_channels$ =1、卷积核大小为7×7的卷积层实现降维和增大感受野后，使用Sigmoid激活函数得到通道信息的注意力权重矩阵M_{SA}。

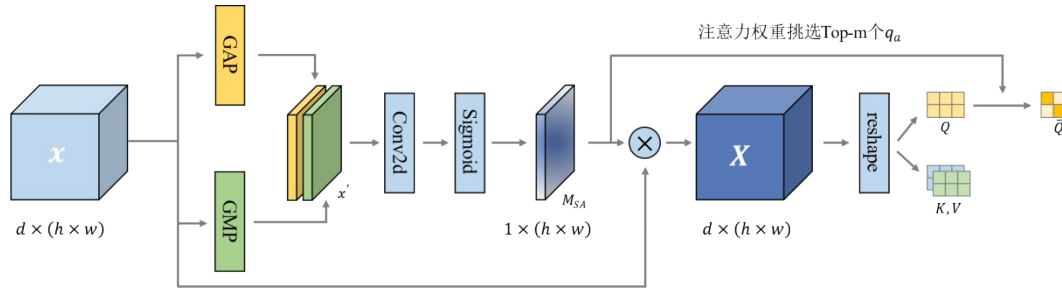


图6 利用类空间注意力算法挑选活跃像素点

如果某个像素位置的通道信息权重M_{SA}越大，则表明此像素点在线性映射为query后，与key的点积结合越有可能查询出信息。

基于此，将通道信息的注意力权重M_{SA}作为q_a的度量方法。对于自注意力层，对输入的每张图像分别进行同样的操作：将重构后的特征图$X \in \mathbb{R}^{(h \times w) \times d}$再次展开为一维向量后，通过不同的参数矩阵$W_q \in \mathbb{R}^{d \times d}$、$W_k \in \mathbb{R}^{d \times d}$、$W_v \in \mathbb{R}^{d \times d}$线性映射为查询矩阵$Q \in \mathbb{R}^{N \times d}$、键矩阵$K \in \mathbb{R}^{N \times d}$、值矩阵$V \in \mathbb{R}^{N \times d}$，将得到的注意力权重$M_{SA}$从大到小进行排序，在$Q$中挑选出其中占主导地位的Top-m个$q_a$（图7(a)），从而实现对所有query的稀疏度评估。根据Zhou等人[22]提出

的策略对m进行定义(式(5)):

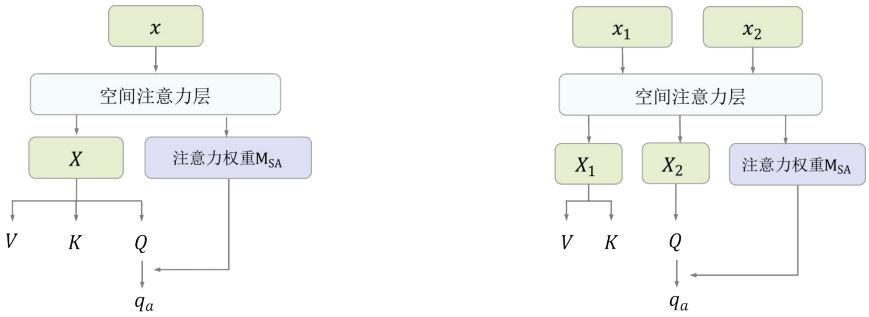
$$m = c \cdot \ln N_{\bar{Q}} \tag{5}$$

其中c为可调超参数。非活跃像素点形成空洞直接由value的平均值填充,最终得到与原始查询矩阵Q大小相同的稀疏矩阵$\bar{Q}$,此时式(5)变为式(6):

$$Attention = softmax\left(\frac{\bar{Q}K^T}{\sqrt{d}}\right)V \tag{6}$$

对于互注意力层,将得到的两个输入向量进行特征重构后,其中一个输出向量X_1线性映射为K和V,另一个输出向量X_2线性映射为Q,同样使用注意力权重进行q_a的挑选。其过程由图7(b)所示。

因为只计算了稀疏度度量下的Top-m个query,理论上每层注意力的时间复杂度降为$O(N \cdot \ln N)$。



(a)自注意力层挑选活跃像素点 (b)互注意力层挑选活跃像素点

图7 自-互注意力层挑选活跃像素点

3.2.2 AttentionRanker——基于排名优化的自-互注意力机制

上文活跃像素点的挑选策略已经确定了每层自注意力和互注意力的运行机制,其流程示意图如图8所示。

对于每一张图像,与Sparse Transformer等启发式稀疏注意力方法不同,AttentionRanker会根据图像的特征自适应地生成不同的空间注意力权重值,每层自注意力和互注意力的输入都会用Top-m思想评估出不同的q_a,计算生成不同的$\bar{Q}K^T$矩阵,从而使得在计算多头注意力时,每张图像上的重要像素点既不会因为注意力头不同而改变,对于每一层的输入又可以自适应选择活跃query从而采取不同的优化策略。

在自注意力层中,其Q_s、K_s、V_s的输入都来自于同一特征向量。在将重构后的特征向量展平并经过不同的线性层转换成表征长度相同的向量后,通过隐藏维度的信息压缩选出空间注意力权重最高的Top-m个q_a,只计算这些q_a和所有key的点积结果,其余的q_l不再进行计算(即不再为value计算权重),而是直接对value取均值作为输出,从而保证输入输出的长度统一。并行计算每个特征图的自注意力,得到带有自身特征关联信息的F_s^A和F_s^B,将其进行特征重构后分别作为互注意力层Q_c和K_c、V_c的输入特征向量,同样进行上述步骤后输出带有相互特征关联信息的F_c^A和F_c^B。将上一层的输出向量作为下一层自-互注意力的输入向量,在N_c次信息传递之后,最终得到融合本张图片邻域信息与待匹配图像信息的输出特征F_{tr}^A和F_{tr}^B。

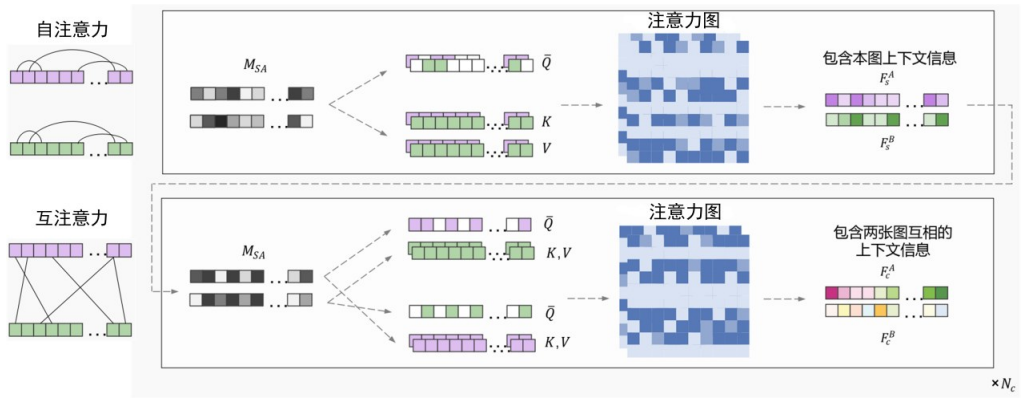


图8 基于排名优化的自-互注意力机制

3.3 无检测器的特征匹配模型

3.3.1 强纹理增强模块

本节介绍在特征金字塔ResNet18-FPN的基础上加入强纹理特征增强模块（Strong Texture Feature Enhancement Module，ST-FEM）后的网络结构。

如图9所示，将ResNet每层特征图的输出表示为$\{C_1, C_2, C_3\}$，自顶向下过程中的每层特征图的输出表示为$\{P_1, P_2, P_3\}$。图9(a)给出了FPN自顶向下过程中P_2级别到P_1级别的融合路径示意图，通过1×1卷积核对C_1进行通道降维，横向连接来自空间域2倍最近邻上采样的特征图P_2和自底向上特征提取过程中相同空间大小的特征图C_1。下文所述的网络结构均为将ST-FEM模块置于1×1卷积前的情况。

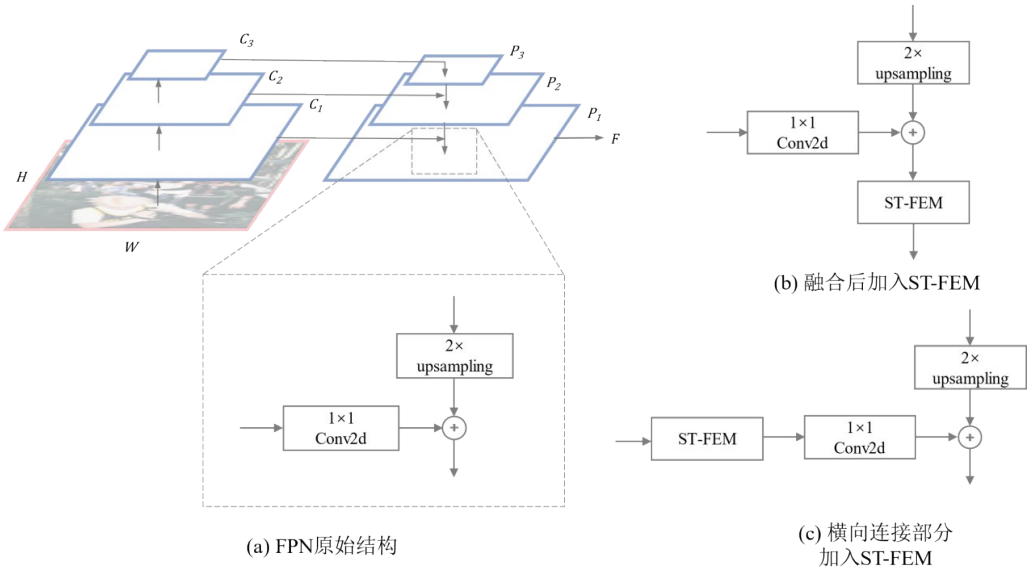


图9 特征金字塔中的ST-FEM模块示意图

2018年Park等人提出的BAM[19-20]中指出在神经网络中，不同的维度所代表的意义不同：对于通道维度而言，其包含的信息更多为特征的抽象表达，而对于空间维度，则拥有更为丰富的特征位置信息。为了使得特征提取网络更加关注于强纹理区域特征，本章将来自于自底向上过程中的除最高层语义的其他尺度特征图（以C_1、C_2为例）进行如下处理：

（1）经过全局最大池化*MaxPool*和全局平均池化*AvgPool*得到不同的语义描述符$\boldsymbol{M} \in \mathbb{R}^{h \times w \times 1}$（式(7)）和$\boldsymbol{A} \in \mathbb{R}^{h \times w \times 1}$（式(8)），即将每个像素点在不同通道上的最大值和平均值表示在空间维度的每个位置中：

$$\boldsymbol{M} = Maxpool(C) \tag{7}$$

$$\boldsymbol{A} = Avgpool(C) \tag{8}$$

（2）将每个像素点在空间维度上进行全局低维嵌入$\{(h \times w) \rightarrow (1 \times 1)\}$，即将$\boldsymbol{M}$和$\boldsymbol{A}$经过全局平均得到整张图的最大值$Avg(\boldsymbol{M})$和平均值$Avg(\boldsymbol{A})$。

（3）将$\boldsymbol{M}$和$Avg(\boldsymbol{M})$相减得到每个像素点与整张图像的差异绝对值描述符$\boldsymbol{M}'$（式(9)），同理得到$\boldsymbol{A}$和$Avg(\boldsymbol{A})$的差异绝对值描述符$\boldsymbol{A}'$（式(10)），绝对值越大则代表这个像素点与周围、与整张图像越不同，即本节所述的强纹理特征区：

$$\boldsymbol{M}' = \left| \boldsymbol{M} - Avg(\boldsymbol{M}) \right| \tag{9}$$

$$\boldsymbol{A}' = \left| \boldsymbol{A} - Avg(\boldsymbol{A}) \right| \tag{10}$$

（4）将带有强纹理特征相对位置的$\boldsymbol{M}'$和$\boldsymbol{A}'$进行拼接，经过卷积核大小为7×7的卷积层$f(\cdot)$和Sigmoid激活函数$\sigma(\cdot)$后，与自底向上过程中提取的特征图C_1、C_2进行融合得到强纹理特征增强的特征图C_1'和C_2'（式(11)）：

$$C' = \sigma\left(f\left(Concat\left[\boldsymbol{M}', \boldsymbol{A}' \right] \right) \right) \tag{11}$$

最后经过1×1卷积形成一个完整的横向连接。整体结构如图10所示。

3.3.2 多尺度自-互注意力融合机制

针对特征金字塔提取的多尺度特征图，采用两种不同的自-互注意力融合设计：

（1）对于粗精度特征图$\tilde{F} \in \mathbb{R}^{60 \times 80 \times 256}$，采用AttentionRanker方法。将$\tilde{F}$展平为一维向量后与绝对正弦-余弦位置编码进行相加融合得到一维特征向量

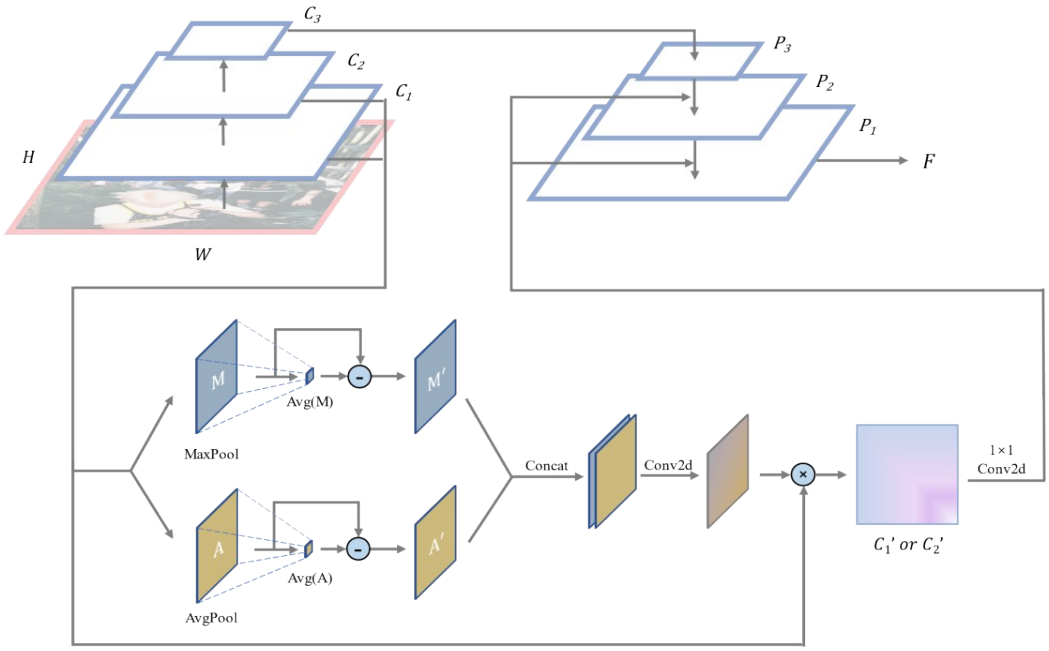


图10 强纹理特征增强模块ST-FEM示意图

$\tilde{F}_{pe} \in \mathbb{R}^{4800 \times 256}$,为了降低计算复杂度进行活跃像素点的挑选,即从原本经过线性映射得到的4800个全部参与注意力点积计算的查询向量q中挑选出Top-m个活跃查询q_a,将$\tilde{F}_{pe}$重新塑形为粗精度特征图大小,通过类空间注意力权重挑选策略在每次送入自注意力层和互注意力层时进行一次挑选。循环N_c次后输出得到充分聚合全局上下文信息的$\tilde{F}_{tr}$。

(2)对于细精度特征图$\hat{F} \in \mathbb{R}^{240 \times 320 \times 256}$,采用Linear Transformer[23]方法进行线性化自-互注意力融合。首先将通过互匹配得分矩阵得到的粗匹配预测在细精度特征图上进行裁剪定位,本文选取窗口大小为5×5的网格作为定位点,然后将$n(n \leqslant 3072)$个5×5的局部窗口展平为一维向量送入线性自-互注意力特征融合模块,即将查询向量和键向量之间的Softmax点积计算转变为基于特征映射的线性注意力计算,以特征映射为$\phi(x) = elu(x) + 1$的相似度函数$sim(Q,K) = \phi(Q) \cdot \phi(K)^T$为注意力计算的核函数近似算法,先一步计算key-value的点积相乘,再与query进行结合,该算法如图11所示。

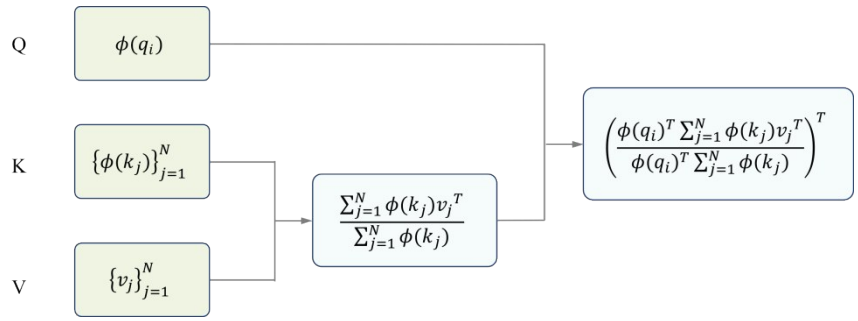


图 11 Linear Transformer的注意力机制

对于粗精度的自-互注意力特征融合步骤,在非稀疏方法下需要进行近五千个点积计算的查询向量中挑选几十个活跃查询可以很大程度上降低计算量,但如果对细精度匹配步骤采用同样的Attention-Ranker稀疏方法,在非常少量的查询向量中挑选活跃像素点意义不大。故本文针对不同尺度的特征图选用了"AttentionRanker +Linear"两种不同的稀疏注意力方法。

3.3.3 损失函数设计

整体算法的搭建包含"由粗到细"的多尺度递进匹配思路,遵循文献[10-11,24]的损失函数设计方案,本文算法最终损失L包括粗精度损失L_c和细精度损失L_f

（如式（12））：

$$L = L_c + L_f \tag{12}$$

（1）粗精度损失 L_c

每个特征都代表原图上的一个像素网格，由于粗精度特征图和细精度特征图是多尺度的，在由粗到细的匹配过程中很可能会存在一对多的匹配结果，因此也难以准确获得粗精度匹配的真值标签。

ScanNet数据集[25]提供相机位姿和深度图，本文采用在训练过程中实时计算出置信矩阵 P_c 作为真值标签的方法：通过衡量两组低分辨率网格中心位置的重投影距离，从而确定互最近邻，即取 $\tilde{F}^A$ 中网格的中心位置，将其投影到与深度图相同的比例，并在数据集中对其深度信息进行索引，基于深度值和已知的相机位姿，将网格中心扭曲到另一张特征图 $\tilde{F}^B$ 上，并将其最近邻作为匹配候选，从 $\tilde{F}^B$ 到 $\tilde{F}^A$ 重复同样的过程。最后基于两组不同方向的最近邻匹配，保留互最近邻的值作为最终粗匹配的真值 M_c^{gt}。

当使用双Softmax方法进行匹配时，将返回的置信矩阵 P_c 上的负对数似然损失作为 L_c（式（13））：

$$L_c = -\frac{1}{\left|M_c^{gt}\right|}\sum_{(i,j)\in M_c^{gt}} \log P_c\left(i, j\right) \tag{13}$$

（2）细精度损失 L_f

细精度级别的自-互注意力融合是在以粗匹配预测为中心的5×5小窗口中进行的。对于每一组粗精度匹配 $\left(i, j\right)$，本文将 $\hat{F}^A$ 网格的中心位置扭曲到 $\hat{F}^B$ 上，计算其与最近邻之间的距离，并对对应匹配点 j' 是否位于细精度特征图 $\hat{F}^B$ 网格的对应5×5窗口进行检查，过滤无法找到对应匹配点的粗匹配预测对，最终获得真值 j'_{gt}。对于细精度特征图 $\hat{F}^A$ 的每个网格中心点，通过计算相应热力图的总方差 $\sigma^2(i)$ 来衡量其不确定性。为了优化具有低不确定性的亚像素级别匹配位置，使用 L_2 损失设计加权细精度损失 L_f（式（14））：

$$L_f = -\frac{1}{\left|M_f\right|}\sum_{(i,j)\in M_f}\frac{1}{\sigma^2(i)}\left\| j' - j'_{gt}\right\|_2 \tag{14}$$

4 实验

4.1 数据集及评价指标

4.1.1 数据集

整体模型基于ScanNet数据集[25]进行了训练、验证和测试。ScanNet数据集是目前室内相机位姿估计任务中使用最广泛且规模最大的室内图像数据集，包含了707个不同大小的真实室内空间类型，根据不同场景的多次RGB-D扫描组成了1513个单目序列，每一个序列都提供了相应的相机内外参数、真实位姿和深度图像。

考虑实验条件，本文在ScanNet数据集的1513个单目序列中使用随机函数Random获得200个编号数。该数据集每一个场景命名方式为其场景编号（0~706）与扫描次数编号（0~3）组成，其中编号为scene0307_00、scene0366_00、scene0412_00、scene0645_00的场景由于解析错误造成数据损坏（其余使用该数据集的算法[25]同样将其做删除处理），故最终构成包含约30万个视图的子数据集ScanNet196。

为保证实验结果的公平性与有效性，本文的所有实验包括其它算法的复现均在ScanNet196上进行。

4.1.2 评价指标

根据本文的算法结构，STEM属于无特征检测器的图像匹配算法，对于此类匹配网络，暂时没有明确的类似匹配分数MS等衡量匹配精度的度量方法，因此本文沿用SuperGlue[11]算法在ScanNet数据集[25]中针对相机位姿估计任务的Pose eatimation AUC评估标准，以旋转和平移的最大角度误差的累积误差曲线的曲线下面积作为评价指标。本文分别取AUC@5°、AUC@10°和AUC@20°的指标进行实验结果分析。

4.2 实验设置及实施细节

4.2.1 实验环境

实验采用PyTorch深度学习框架下的Python 3.8语言进行编程，在Ubuntu18.04操作系统下使用3块GPUs (NVIDIA RTX A5000) 对模型进行训练。实验环境具体配置如表1所示。

表1 实验环境配置

项目	名称/版本
Operating System	Linux (Ubuntu18.04)
IDE	Jupyter Notebook
CPU	AMD EPYC 7543*3, 45核
GPU	NVIDIA RTX A5000*3
编程语言	Python 3.8
框架	PyTorch
计算机开源视觉库	OpenCV

4.2.2 训练细节

使用初始学习率为 6×10^{-3},批量大小(Batch size)为64的Adam优化器[26]对模型进行70个周期的训练。学习率的调整策略为线性缩放规则(Linear Scaling Rule):先线性预热4800次迭代(iteration),从第3个周期开始,每3个周期学习率衰减0.5。每个周期训练结束后,自动保存验证结果,最终保存各项指标最优的5个结果。

整个模型采用随机初始化权值进行端到端训练。基于排名优化的自-互注意力方法在粗精度阶段循环4次,其中采样超参数c设置为5,即每次挑选Top-45个活跃query;细精度阶段使用基准网络LoFTR的Linear Transformer方法循环1次,即 $N_c = 4$, $N_f = 1$。设置置信度阈值 θ_c 为0.2,窗口大小5×5。粗精度特征图 $\tilde{F}$ 和细精度特征图 $\hat{F}$ 的大小分别是原图的1/8和1/2。

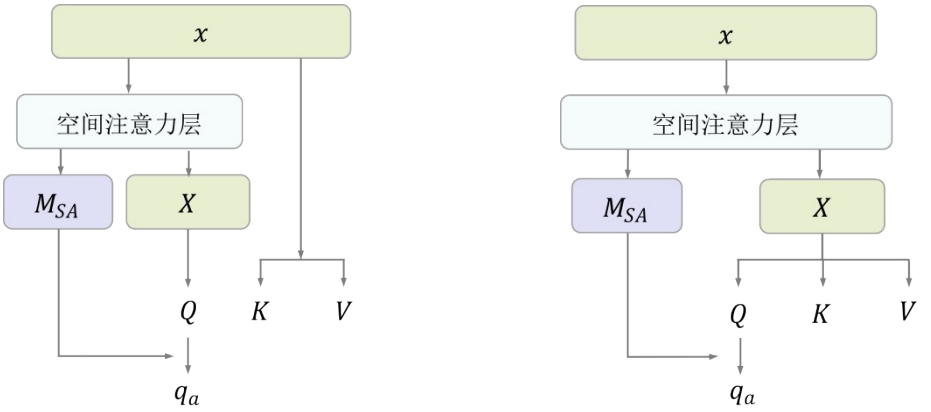
4.3 实验结果分析

本节以2021年图像匹配任务榜首的LoFTR[10]作为基准网络进行对比试验,由于实验环境及配置等因素限制,仅在ScanNet数据集中随机挑选196个场景进行训练,并在1500对图像上进行验证与测试。

(1)消融实验

上文提到的方法是将输入特征图 x 通过类空间注意力机制进行重构后,通过不同的参数矩阵将其线性映射为查询矩阵 Q、键矩阵 K 和值矩阵 V,然后利用注意力权重 M_{SA} 对查询向量query进行稀疏度评估。为了探究此处特征重构对key-value键值对在进行自-互注意力信息融合是否也有一定的积极作用,故设计三个消融实验,并以实验1、2、3来代指。

实验1为不进行类空间注意力挑选活跃像素点的实验情况。实验2直接将输入特征图 x 进行线性映射得到key-value键值对,特征重构后的输出向量 X 映射为query并进行后续活跃像素点的挑选。实验3则是query和key-value都经过特征重构的实验情况。以自注意力层为例,实验2、3的处理方式分别如图12 (a)、(b)所示。



(a)输入特征 x 直接映射为K、V　　(b)特征重构的输出特征 X 映射为Q、K、V

图12 两种不同的特征映射方式

从表2的结果可以看出,同时对比实验1、2、3,仅对query进行特征重构和活跃像素点挑选,位姿估计精确度在各阈值下仅有少量的提升,而如果在线性映射为key-value之前也进行了隐藏维度的信息压缩,其Pose estimation AUC则会在5°、10°、20°阈值下在前者(实验2)的基础上再提升0.47%,1.75%和1.06%,说明输入特征 x 的特征重构可以加强整体自-互注意力信息融合阶段的特征信息表达。

表2 特征映射消融实验结果

序号	是否特征重构		Pose estimation AUC(%)		
	key-value	query	@5°	@10°	@20°
1			14.73	32.53	50.36
2		√	14.98(↑0.25)	33.20(↑0.67)	50.69(↑0.33)
3	√	√	**15.45(↑0.72)**	**34.95(↑2.42)**	**51.75(↑1.39)**

（2）注意力方法的对比实验

在验证集上的进行自-互注意力模块的对比实验，由于普通Transformer空间复杂度过高，表3中第一行数据为使用6块GPU进行训练、验证得到的结果。为保证结果精确性，计算最优本地结果的平均值并保留两位小数。实验主要对比普通Transformer方法以及两种不同的稀疏注意力算法在LoFTR基准网络上的室内位姿估计精度。

表3 自-互注意力模块的对比实验

类别	复杂度（每层）	Pose estimation AUC（%）			模型时耗
		@5°	@10°	@20°	
Transformer	$O(N^2 \cdot d)$	14.56	32.08	50.18	350ms
Linear Transformer	$O(N \cdot d^2)$	14.73	32.53	50.36	202ms
AttentionRanker	$O(N \cdot lnN \cdot d)$	**15.45**	**34.95**	**51.75**	**184ms**

在特征向量长度$N = 4800$，表征维度$d = 256$的情况下，AttentionRanker方法在位姿估计精度（AUC@5°、10°、20°）上比普通Transformer算法分别高0.89%、2.87%、2.37%。同时对比LoFTR文章中提到的线性稀疏注意力算法Linear Transformer，在输入两张图片进行位姿估计的整体耗时上也比前者快18ms。这说明AttentionRanker算法不仅在理论层面降低了时间复杂度，在执行实际的室内姿态估计任务时，也能消耗更少的时间。

（3）整体结果分析

根据表4结果显示，在ScanNet196数据集下，将AttentionRanker应用到室内位姿估计任务后在阈值为10°和20°的情况下表现出了最好的效果，分别达到了34.95%和51.75%。与曾经基于特征检测器的图像匹配最优算法SuperPoint和SuperGlue相比，本文算法能够很大程度提高位姿估计精度，并且仅在阈值为5°时略逊色于2022年的四叉树注意力算法（LoFTR-QuadTreeB）。本文的方法在进一步降低计算复杂度的同时，可以维持甚至优于当前室内位姿估计的最优算法，这说明自适应稀疏自-互注意力机制在轻量化Transformer类室内位姿估计任务的同时，也能更好地感知图像中的相关信息。

表4 在ScanNet196数据集上的室内位姿估计结果

类别	Pose estimation AUC（%）		
	@5°	@10°	@20°
SuperPoint+NN	5.63	16.54	23.82
SuperPoint+SuperGlue	11.21	29.42	46.37
LoFTR-QuadTreeB	**15.76**	34.23	51.33
LoFTR- AttentionRanker	15.45	**34.95**	**51.75**

5 结论

本文对现有的图像匹配算法展开了研究，针对在匹配融合阶段引入Transformer带来的计算复杂度高这一问题，设计了面向计算机视觉任务的基于排名优化的自-互注意力机制AttentionRanker。该算法通过对位置编码后的一维输入特征图进行重塑形，利用类空间注意力机制挑选少量活跃像素点，成功地将点积注意力的时间复杂度从二次降为近线性。实验结果表明，采用了AttentionRanker稀疏方法的网络在前向推理时耗时比基准网络快18ms，且其Pose estimation AUC@5°/10°/20°相较于Linear Transformer方法分别提升了0.72%、2.42%、1.39%。

参考文献（References）：

[1] Marr D. Vision：A Computational Investigation into the Human Representation and Processing of Visual Information[M]. San Francisco：W. H. Freeman，1982.

[2] Ma J，Ma Y，Li C. Infrared and visible image fusion methods and applications：a survey[J]. Information Fusion，2019，45：153-178.

[3] Radke R J，Andra S，Al-Kofahi O，et al. Image change detection algorithms：a systematic survey[J]. IEEE Transactions on Image Processing，2005，14(3)：294-307.

[4] Zheng L，Yang Y，Tian Q. SIFT meets CNN：a decade sur-

vey of instance retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1224-1244.

[5] Fan B, Kong Q, Wang X, et al. A performance evaluation of local features for image-based 3D reconstruction[J]. IEEE Transactions on Image Processing, 2019, 28(10): 4774-4789.

[6] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendon-Mancha J M. Visual simultaneous localization and mapping: a survey [J]. Artificial Intelligence Review, 2015, 43:55-81.

[7] Liu C, Yuen J, Torralba A. SIFT flow: dense correspondence across scenes and its applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5):978-994.

[8] Rocco I, Cimpoi M, Arandjelović R, et al. Neighbourhood consensus networks[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018:1658-1669.

[9] Li X, Han K, Li S, et al. Dual-resolution correspondence networks[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 17346-17357.

[10] Sun J, Shen Z, Wang Y, et al. LoFTR: detector-free local feature matching with transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:8918-8927.

[11] Sarlin P E, Detone D, Malisiewicz T, et al. SuperGlue: learning feature matching with graph neural networks[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020:4937-4946.

[12] Sinkhorn R, Knopp P. Concerning nonnegative matrices and doubly stochastic matrices[J]. Pacific Journal of Mathematics, 1967, 21:343-348.

[13] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013:2292-2300.

[14] Zaheer M, Guruganesh G, Dubey K A, et al. Big bird: transformers for longer sequences[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020:17283-17297.

[15] Child R, Gray S, Radford A, et al. Generating long sequences with sparse transformers[DB/OL]. arXiv:1904.10509, 2019.

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000 - 6010.

[17] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision, 2020:213-229.

[18] Zhou H, Zhang S, Peng J, et al. Informer: beyond efficient transformer for long sequence time-series forecasting [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021:11106-11115.

[19] Park J, Woo S, Lee J Y, et al. BAM: bottleneck attention module[DB/OL]. arXiv:1807.06514,2018.

[20] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision, 2018:3-19.

[21] Lin M, Chen Q, Yan S. Network in network[DB/OL]. arXiv:1312.4400, 2014.

[22] Zhou H, Zhang S, Peng J, et al. Informer: beyond efficient transformer for long sequence time-series forecasting [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021:11106-11115.

[23] Katharopoulos A, Vyas A, Pappas N, et al. Transformers are RNNs: fast autoregressive transformers with linear attention[C]//Proceedings of the 37th International Conference on Machine Learning, 2020:5156-5165.

[24] Wang Q, Zhou X, Hariharan B, et al. Learning feature descriptors using camera pose supervision [C]//European Conference on Computer Vision (ECCV), 2020: 757 - 774.

[25] Dai A, Chang A X, Savva M, et al. Scannet: richly-annotated 3d reconstructions of indoor scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2432-2443.

[26] Kingma D P, Ba J. Adam: a method for stochastic optimization[DB/OL]. arXiv:1412.6980, 2014.

编辑:赵志军,龙学锋