引用格式：彭宏，侯小刚，曾凡璐，吴萌.融合金字塔和注意力机制的文物子图检索模型[J].中国传媒大学学报（自然科学版），2024，31(02)：19-26.
文章编号：1673-4793(2024)02-0019-08

融合金字塔和注意力机制的文物子图检索模型

彭宏[1]，侯小刚[2,3*]，曾凡璐[3]，吴萌[4]
（1.文化和旅游部民族民间文艺发展中心，北京 100007；2.中国国家博物馆，北京 100006；
3.北京邮电大学人工智能学院，北京 100876；4.北京故宫博物院，北京 100006）

摘要：随着中国文化研究工作的深入以及数字化文物采集技术的发展，文化资源数据和文化数字内容的数量也随之增长，如何对文化数据进行有效存储、管理以及检索成为一项重要的工作。针对文物图像数据检索任务中因尺度变化和特征选择造成检索精度不高的问题，提出了一种融合折叠多空洞金字塔池化和注意力机制的文物子图检索模型。模型为提高不同尺度的文物子图检索精度，通过在图像特征提取模块使用优化后的折叠多空洞金字塔池化提取图像的多尺度信息；为避免密集局部特征和无关特征影响检索准确率，使用注意力机制对局部特征进行关键特征选择。最后在所构建的文物数据集上进行了消融实验和性能对比实验，实验结果取得了良好的效果，mAP达到85.3%。

关键词：子图检索；空洞金字塔；注意力机制；特征选择；图像检索
中图分类号：TP319.56　**文献标识码**：A

A cultural relic sub-image retrieval algorithm based on folded multi-hollow pyramid pooling and attention mechanism

PENG Hong[1], HOU Xiaogang[2,3*], ZENG Fanlu[3], WU Meng[4]
(1.Center for Ethnic and Folk Literature and Art Development, Ministry of Culture and Tourism, PRC, Beijing 100007, China; 2. National Museum of China, Beijing 100006, China; 3. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; 4. The Palace Museum in Beijing, Beijing 100006, China)

Abstract: With the deepening of research on Chinese culture and the development of digital cultural relics collection technology, the amount of cultural resource data and cultural digital content has also increased, so how to store, manage and retrieve cultural data has become an important task. In order to solve the problem of low retrieval accuracy caused by scale change and feature selection in cultural relic image retrieval tasks, in this paper a cultural relic sub-image retrieval algorithm based on folded multi-hollow pyramid pooling and attention mechanism (FMHPPA) was proposed. In order to solve the problem of scale change in sub-image retrieval, FMHPPA model extracted multi-scale information from image feature extraction module by optimizing folded multi-hollow pyramid pooling. In order to avoid the impact of dense local features and irrelevant features on retrieval performance and accuracy, FMHPPAM model used attention mechanism to select key features for local features. The model ablation experiment and performance comparison experiment were carried out on the

基金项目：国家重点研发计划项目（2022YFF0904304）；内蒙古自治区科技计划（2023YFSW0021）
作者简介（*为通讯作者）：彭宏（1972—），男，高级工程师，主要研究领域为文化资源数字化。Email：466985365@qq.com；侯小刚（1985—），男，博士，工程师，主要从事文化计算、计算机视觉相关研究。Email:houxiaogang05@bupt.cn

constructed sub-image dataset, and the experimental results achieved better results as the mAP reached 85.3%.

Keywords: sub-image retrieval; hollow pyramid; attention mechanism; feature selection; image retrieval

1 引言

随着文化资源数据和文化数字内容数量的快速增长,如何对文化资源的数据和元数据进行有效管理,并建立灵活多样的检索体系成为了一个重要的问题[1-2]。由于文物采集需要对文物的不同视角、不同部位进行拍摄,从而会产生多张描述同一文物但又存在差别的不同图像。子图检索的目的是通过某个文物的图像或部分图像,检索到包含该文物的图像,实现子图检索对于构建文化资源数据和文化数字内容的检索体系具有重要意义。

子图检索的主要任务是通过某物体的图像或部分图像检索包含该物体的图像,从本质上来说和细粒度图像检索任务一致。传统的图像检索主要是关注图像整体的相似性,用内容、颜色以及形状进行检索。而相比于传统图像检索,子图检索更加关注图像的局部特征,细粒度图像通常属于同一大类下的不同子类,不同类之间在外观上有着相似性,而相同类之间在姿态上千变万化,导致数据呈现类间差异性大、类内差异性小的现象,加之数据库中的图像中存在背景干扰、视角变化、姿态变化、光照变化等影响,使得子图检索相比于传统图像检索具有更大的挑战。

2 国内外研究现状

子图检索本质上是细粒度图像检索,而细粒度图像检索的实现方法主要可以分为三类:强监督方法、无监督方法和弱监督方法。

基于无监督信息的细粒度图像检索在训练模型时,不仅没有精细的物体标注信息,而且图像级别标注信息都无从获取[3]。Wei等提出使用预训练的CNN(Convolutional Neural Networks)模型[4],通过将卷积层特征转换为自适应加权的特征向量来表示每个图像,并使用这些向量进行图像检索。Bai等提出将对抗性训练和注意力机制结合解决细粒度图像检索问题[5],该框架将生成器和鉴别器重新设计,以确保生成器检索相似的图像,而鉴别器选择不相似的图像,并为生成器创建对抗性奖励,通过对抗性奖励检索机制进行极小极大对抗,直到鉴别器无法判断检索到图像序列是否与查询相似。

强监督方法需要使用详细的标注信息,例如标记出图像中每个物体的位置和类别,这些标注信息的获取通常需要大量的时间和人力成本。而弱监督方法则只需要使用相对简单的标注信息,例如图像级别的标签或者部分标注信息。Zheng等提出了一个新的中心化排序损失函数用于提升训练速度[6],并结合弱监督特征提取方法,以自上而下的显著性分割物体轮廓,将轮廓集成到CNN特征图中,以精确地提取目标对象内的特征,避免背景噪声对检索产生影响。Noh等使用经过图像级标签训练后的卷积神经网络获得的局部特征[7],结合注意力机制对关键点进行选择实现检索过程,并且该研究引入了一个新的大规模数据集,该数据集称为Google Landmarks数据集,包含背景杂波、部分遮挡、多个地标、可变尺度的对象等问题。Phan等将全局和局部特征与Vision Transformers和多空洞卷积相结合实现检索[8],在Vision Transformers编码器层的输出上添加一个空洞卷积模拟图像金字塔解决处理不同图像实例之间的比例变化,使用类注意力来聚合从多空洞卷积层输出的embeddings,以获得全局和局部特征。Cao等将全局和局部特征统一到一个单独的深度模型中[9],从而通过有效的特征提取实现准确的检索。

相比于弱监督和无监督方法,强监督方法在获得图像标注信息上需要花费大量成本,但同时也获得了比无监督和弱监督方法更好的效果。一般情况下,强监督方法不仅仅需要图像级别的标签,还需要在图像上提供标注框、标注点和对应的标签信息。Mohedano等提出了一种简单的强监督细粒度检索方法[10],该方法基于使用词包对CNN的卷积特征进行编码,将激活图的每个局部特征分配给视觉单词产生分配图,然后使用分配图进行重排序,获得用于查询图像的对象定位。Salvador等提出从物体检测Faster R-CNN中得到的区域性的局部特征用于实例检索的适用性[11],利用区域建议网络(Region Proposal Network, RPN)学习到的区域对象提案及其相关的CNN特征来构建特征集,再进行相似度计算和排序得到目标图像。Teichmann等对Google Landmarks数据集拓展提出了一个新的地标边界框数据集来填补缺乏物体边界框数据集的空白[12],该研究引入了一种新的区域聚集选择性匹配核(R-ASMK),可以有效地将来自检测区域的信息组合到改进的整体图像表示中。

上述算法都是应用于一般图像检索任务或细粒度图像检索任务中，而针对文物子图检索任务的算法相对较少。本文针对文物图像数据检索任务中因尺度变化和特征选择造成检索精度不高的问题，提出了一种融合折叠多空洞金字塔池化和注意力机制的文物子图检索模型。

3 融合FMHPPA的子图检索模型

本文以文物图像数据为研究对象，针对子图检索任务中因目标尺度变化和特征选择造成检索精度不高的问题，提出了一种融合折叠多空洞金字塔池化和注意力(Folding Multi-Hole Pyramid Pooling and Attention，FMHPPA)机制的子图检索模型。为解决子图检索存在的尺度变化问题，FMHPPA模型在图像特征提取模块使用优化后的折叠多空洞金字塔池化提取图像的多尺度信息。为避免密集局部特征和无关特征影响检索准确率，FMHPPA模型使用注意力机制对局部特征进行关键特征选择。子图检索算法框架如图1所示，图像库中的图像和查询图像经过折叠多空洞金字塔池化提取多尺度信息，然后使用注意力机制去除无关特征，最后通过计算图像库中图像和待查询图像的局部特征的相似度，得到相似度最高的top-k图像即检索结果。

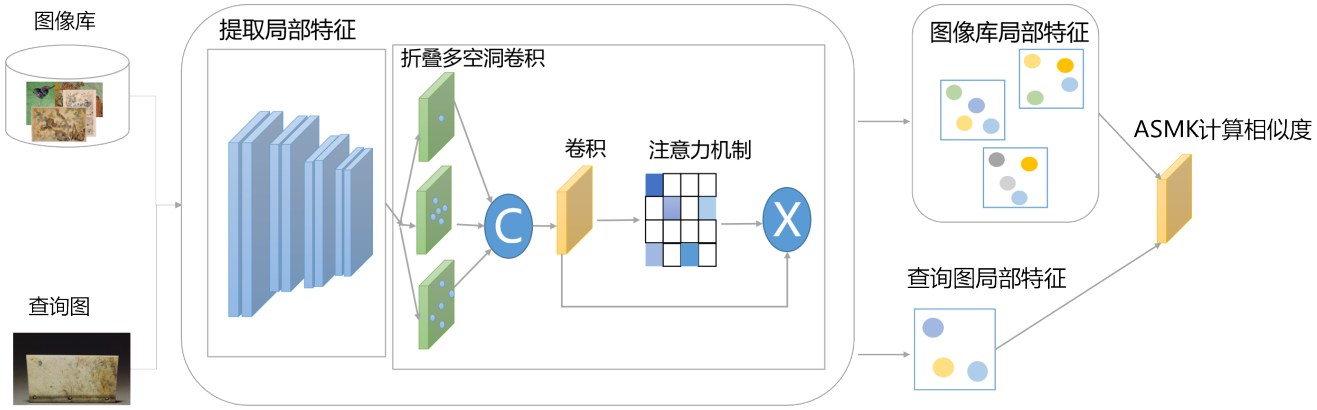

图1 子图检索算法框架图

3.1 多尺度密集特征提取

全局特征可以提供图像的整体信息，帮助识别不同类别之间的差异；而局部特征则可以捕获物体的细节特征，提高识别的准确性。由于子图检索和图像的局部区域有关，因此可以利用CNN中卷积和池化操作的局部感知特性，提取到更多的局部信息。同时为了解决查询图像的尺度变化问题，本文使用CNN模型中的ResNet[13]作为Backbone，并在其conv4_x模块的输出后添加优化后的折叠多空洞金字塔池化模块提取多尺度信息。

多空洞金字塔池化ASPP(Atrous Spatial Pyramid Pooling)[14]是结合空间金字塔池化SPP(Spatial Pyramid Pooling)和空洞卷积的一种池化方式，能够有效地捕获多尺度信息。该模型采用了多个扩张率的空洞卷积，使模型在多尺度物体上的表现更好。ASPP示意图如图2所示，采用了四种不同采样率的空洞卷积来获取更大的感受野从而捕捉多尺度信息，同时这四种空洞卷积的输出特征图的大小是一致的，最终可以将其合并。

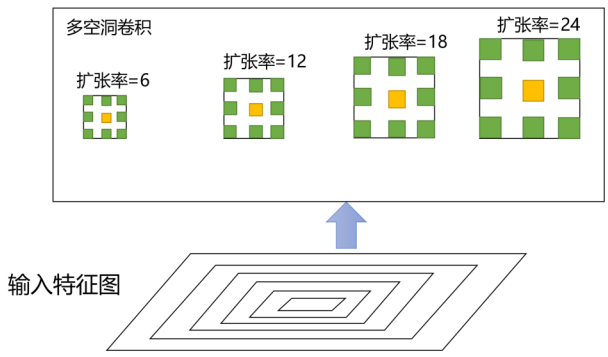

图2 空洞金字塔池化示意图

ASPP使用多个平行的具有不同膨胀率的空洞卷积，空洞卷积虽然可以扩大感受野和捕获多尺度上下文信息，但是由于空洞卷积在卷积核内采样多个孤立的点，可能会导致局部信息的丢失，而且由于不同位置空洞采样的内容可能没有相关性，可能会导致远距离获取的信息没有被充分利用，不能得到稳定的特征。参考Zhao等[15]的实现，本文引入简单的“展开”和“还原”操作来解决这个问题，不仅可以扩大感受野，而且可以采样更大的区域避免采样孤立点，从而解决

局部信息丢失和相关性丢失问题,在本文中称为折叠多空洞金字塔池化。具体实现如图3所示。

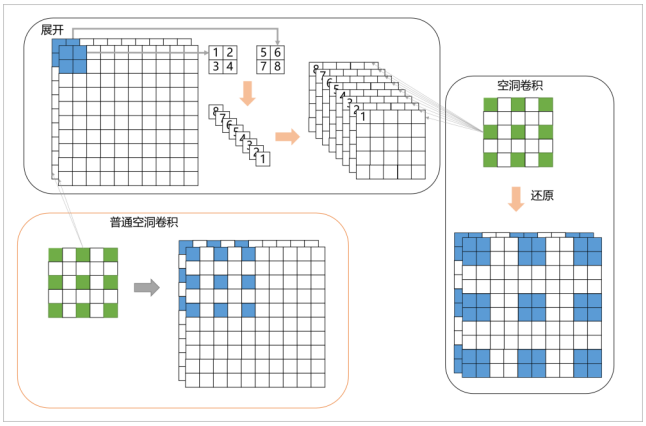

图3 展开还原空洞卷积示意图

假设M表示大小为$H\times W\times C$(H为特征图的长,为特征图的宽,C为通道数)的特征图,使用一个大小为2×2、步长为2的滑动窗口在M上滑动,并在此窗口上使用不同扩张率的尺寸为$K\times K$的卷积核构建多空洞卷积。在图3中,特征图M的大小为$H\times W\times C$=10×10×2,首先把特征图展开,即在特征图M上滑动窗口所采样的$2\times 2\times C$=2×2×2区域展开,那么展开后特征图M的大小变为$H/2\times W/2\times 4C$=5×5×8,再使用尺寸为K=3、扩张率为2的空洞卷积核分别对$4C$=8个特征矩阵做卷积运算,将卷积后的结果再还原。由于在更多的部位上做了卷积运算,相比于普通空洞卷积的结果,展开还原空洞卷积的结果聚合到了更多的信息,同时局部点之间的联系也被保留了下来,提取到了更稳定的局部特征。

图3仅展示了扩张率为2的展开还原空洞卷积的结果。为了获取多尺度信息,本文将ResNet50第四层的输出特征分别使用扩张率分别为2、4、6的空洞卷积核做卷积运算,还包括一个1×1的卷积层和一个全局平均池化层,五个分支的输出沿着channels方向进行拼接,然后再通过一个1×1的卷积将通道数目减少并融合信息,得到的特征矩阵被输出到注意力模块做进一步的处理。通过此步骤后获取的局部特征不仅获得了CNN本身存在的平移不变性,而且具有较好的稳定性和尺度不变性。

研究表明特征提取模块和注意力模块一起训练得到的模型效果较差,因此本文采用两阶段训练的策略,在训练多尺度特征提取模块的时候去掉注意力模块,同时放在分类任务中进行微调,仅需要图像级别的标签即可完成训练。多尺度特征提取模块的目标便是学习到一个特征映射函数$\varphi(x;\theta)$,其中θ是参数,x表示为一张图像,网络的输出y是对Q个类别的预测概率,计算如式(1)所示:

$$y=\varphi(x;\theta) \tag{1}$$

本文使用交叉熵损失进行训练,损失函数如式(2)所示:

$$L=-y^{*}\cdot\log\left(\frac{\exp(y)}{\boldsymbol{I}^{T}\exp(y)}\right) \tag{2}$$

其中,y^{*}是ground-truth的one-hot编码,$\boldsymbol{I}$为单位矩阵。

3.2 关键特征选择

经过卷积、池化等操作产生的局部特征不仅比较密集,而且会有许多无关局部特征存在,会产生计算复杂度高且影响检索准确率的问题,因此采取特征选择手段来选择关键局部特征是必要的。特征选择就是从给定的特征集中选择出相关特征子集的过程,由于注意力机制的生成过程和特征选择过程极为相似,都是在高维数据集里选择部分数据,因此被应用到了特征选择领域内。

本文中注意力模块的主要目的是产生一个分数矩阵表示各个局部特征的相关性分数,从而为特征的选择提供依据。由于采用两阶段训练的策略,本文在训练注意力模块的时候,将折叠多空洞金字塔池化模块替换为注意力模块,然后同样放在分类任务中进行训练,在训练分类器的结果中得到注意力模块参数,且仅需要图像级别的标签即可完成训练。

本文将某个图像的N个d维的特征向量表示为$f_i\in\mathbb{R}^{d}, i=1,\cdots,N$,目标是学习一个评分函数对每个特征向量输出一个分数,以表示该局部特征的相关性,该函数表示为$\phi=(f_i;\theta), i=1,\cdots,N$,其中$\theta$是该函数的参数。为了避免学习到的局部特征相关性分数是负数,本文引入了sofplus激活函数限制$\phi(\cdot)$的输出为正数,对于分类网络的输出y,可以通过如式(3)计算:

$$y=W\cdot\left(\sum_{i=1}^{N}\phi(f_i;\theta)\cdot f_i\right) \tag{3}$$

其中,$W=\mathbb{R}^{M\times d}$表示最后全连接层预测Q个类别的权重。在训练过程中,同样使用交叉熵损失函数和反向传播机制对参数进行调整,优化目标通过式(2)计算得到。

经过注意力机制进行特征选择后,图像的一些关键局部特征会被保留下来用于后续的相似度计算。为了使得选择的关键点可视化,将查询图像和目标图

像选取的关键特征和匹配关系可视化，具体的展示如图4所示。图中黑色圆圈标出的是注意力模块最终选取的分数最高的关键局部特征（为了方便可视化，避免匹配过多导致蓝色线条遮挡，此处限制了选取关键特征的数目），蓝色线条是左右两张图之间的局部特征相匹配的情况。(a)、(b)、(c)均是完整图像和子图之间的局部特征匹配情况，(d)是同一个物体在不同视角下拍摄的图像之间的局部特征匹配情况。

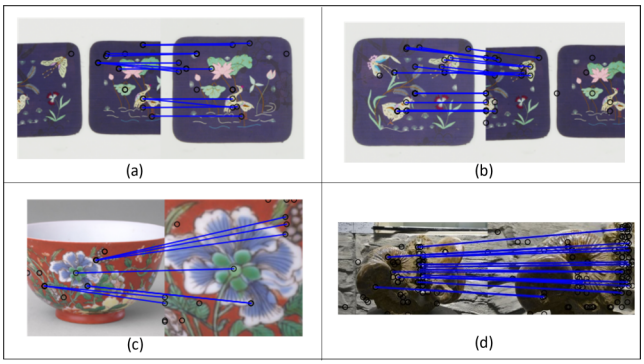

图4 关键局部特征示意图

3.3 局部特征编码及相似度计算

结合注意力模块输出的相关性分数矩阵，对查询图像和图像库中的每张图像都提取n个分数最高的d维局部特征向量，表示为$X=\{x_1,\cdots,x_n\}$，然后使用ASMK方法[11]将得到的局部特征进一步编码得到一个紧凑的特征描述符。

首先对n个局部特征向量进行k-means聚类：q: $\mathbb{R}^d \to C \subset \mathbb{R}^d$，$X \to q(x)$。其中$C=\{c_1,\cdots c_k\}$是由局部特征聚类后的$k$个聚类中心组成的集合，其中的每个元素称为视觉词。$x_c=\{x \in X:q(x)=C\}$是$X$中被分配到视觉词$C$的局部特征的集合。基于局部特征计算图像相似度的一般形式如式(4)所示：

$$SIM(x,y)=\delta(x)\delta(y)\sum_{c\in C}W_cL(x_c,y_c) \tag{4}$$

其中L被称为匹配核，W_c是依赖于视觉词C的常数，$\delta(\cdot)$为标准化系数，采用式(5)计算得到。

$$\delta(X)=\left(\sum_{c\in C}w_cL(x_c,y_c)\right)^{-\frac{1}{2}} \tag{5}$$

通过(5)式，可以使得一张图像的自相似性$SIM(x,x)=1$。

如图5，对于匹配核L，计算分配给视觉词C的两张图像的局部特征间的相似度，而两张图像的相似度通过各个视觉词的匹配核乘以系数再累加。

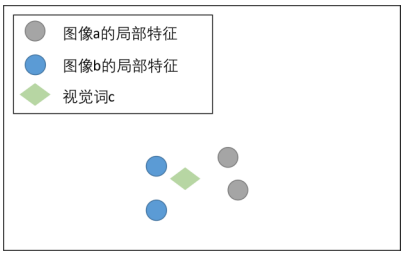

图5 匹配核示意图

匹配核主要分为两种，基于聚合的方法和基于匹配的方法。基于聚合的方法是将两张图像的局部特征描述符先聚合，再计算相似度；基于匹配的方法是将两张图像的局部特征交叉匹配，再把局部特征间的相似度累加。相比非聚合核，聚合核先聚合描述子的操作能够获得更加紧凑的向量表征，使得时空资源开销更小。ASMK是基于聚合的匹配核，计算公式如式(6)所示：

$$L(x_c,y_c)=\sigma_\alpha\left(\hat{V}(x_c)^T\hat{V}(y_c)\right) \tag{6}$$

其中，$\sigma_\alpha(\cdot)$是一个选择性函数，目的是设置一个阈值排除掉匹配度较低的视觉词，计算如(7)式所示：

$$\sigma_\alpha(u)=\begin{cases}sign(u)|u|^\alpha, & if\ u>\tau\\ 0, & otherwise\end{cases} \tag{7}$$

其中，α是一个超参数，α越大说明选择性越强，即只有匹配的描述子的相似度足够大时才能对匹配核有比较大的贡献，从而避免错误的描述子对匹配核的影响，本文设定$\alpha=3$。

$V(x_C)$是将图像局部特征集合成视觉词C的聚合特征，$\hat{V}(x_C)$是$V(x_C)$归一化后的值，计算公式如(8)式所示：

$$\begin{aligned}V(x_C)&=\sum_{x\in X_C}(x-q(x))\\ \hat{V}(x_C)&=V(x_C)/\|V(x_C)\|\end{aligned} \tag{8}$$

ASMK方法计算图像相似度的流程如下：

(1)训练码本并聚类。在查询图像和图像库中的图像计算相似度时，如果每计算两张图像的相似度就需要进行一次聚类，那么时间开销是很大的，所以可事先用训练集的描述子先进行聚类，然后保存聚类中心，这些聚类中心的集合就是所谓的码本。也就是将聚类这一步提前抽取出来，训练出一个足够大的码本，满足大量图像的相似度计算，避免多次聚类浪费时间。

(2)计算每个簇的相似度。两张图像中分配到同一个视觉词的局部特征为一个簇，通过计算局部特征与聚类中心的残差的和作为该图像在该簇上聚类中心的距离。

(3)去除语义相似度小的簇。两个描述子即使属于同一个簇,如果它们的内积还是很小的话,它们语义上可能还是没有太大关联的,所以如果内积小于某个阈值的话就直接抛弃掉,如果大于某个阈值,再对这个内积求α次方,也就是希望内积足够大时才会对图像整体的相似度有比较大的贡献。

(4)计算相似度。将每个簇的相似度加起来并归一化,然后计算相似度。

4 实验结果与分析

4.1 数据集构建

本文实现的子图检索模型的主要研究对象是文物图像,构建的数据集的数据为文物图像,包括青铜器、玉器、陶瓷、石器、金银器等。由于本文所构建的子图检索模型在训练过程中仅仅需要图像级别的标注,不需要对象级或物体级的标注,因此构建的训练集以文物对象为类别,单个文物的多张不同视角的图像为同一个标签。本文所构建的子图检索数据集的训练集部分图像如图6所示,可以看到同一个文物在不同的视角下具有不同的图像,共收集到1986张图像。本文按照7:3的比例将收集到的图像划分为了训练集和验证集,其中训练集为1391张,验证集为595张,测试集为652张。

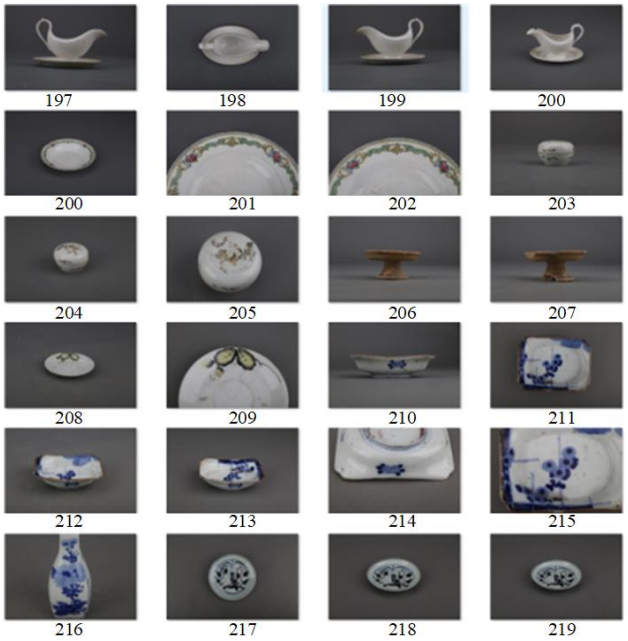

图6 子图检索训练集部分图像示意图

为了验证子图检索模型的效果,本文构建了子图检索测试集。测试集的查询图像主要分为两种:一种是通过对原图像使用人工裁剪的方式获得其中物体的部分图像;另一种是同一个文物的不同视角下获取的图像,以验证本文所提算法在尺度变换、视角变换上的有效性。在652张测试集图像中,部分查询图像如图7所示。

图7 测试集部分查询结果图像示意图

4.2 评价指标

在图像检索中,评价指标用于评估检索结果的质量和性能,以确定算法的有效性和实用性。在图像检索任务中,常见的评价指标包括PR曲线、准确率(Acc)、查准率(Precision Ratio)、F-Score、图像检索精度(AP)、查全率(Recall Ratio)、图像平均检索精度(mAP)等,本文在子图检索的实验中使用了图像平均检索精度(mAP)作为衡量模型性能的指标。

mAP为图像平均检索精度,是一种常用于评估图像检索性能的指标。假设图像库中与输入图像相似的图像数为N_s,在T次排序为前l的检索结果中共检索到了k_l张相似图像,其中正确的检索到了n_l张相似图像,则mAP=$(n_1/1+n_2/2+\cdots+n_l/k_l)/T$。

4.3 实验设置

本文使用2.2.0的Tensorflow搭建子图检索模型,使用在ImageNet上预训练的ResNet50作为特征提取

的Backbone。为了增强局部描述符在本数据集上的相关性，先对ResNet50的前四层加上折叠多空洞金字塔池化模块微调，使用交叉熵损失用于图像分类，最初对输入图像进行中心裁剪以生成方形图像，并将其重新缩放为250×250，然后随机使用224×224裁剪进行训练。

之后对注意力模块进行训练，固定住信息提取模块以产生固定的局部特征，同样采取交叉熵损失进行训练，为了提高注意力模型对物体大小的鲁棒性，采取多尺度训练的策略，最初对输入图像进行中心裁剪以生成方形图像，并将其重新缩放为900×900，然后随机使用720×720裁剪，每隔10 epoch，用一个因数$\xi \leqslant 1$来进行随机尺度变换，再输入到网络中训练。

使用动量为0.9的SGD优化器进行训练，初始学习率为0.05，权重衰减因子设置为0.0001，并采用余弦学习速率衰减策略。

选取注意力分数矩阵中分数最高的1000个局部特征作为ASMK的输入，ASMK选择性函数$\delta_{\alpha}(\cdot)$中设置$\alpha = 3$。

4.4 实验结果分析

（1）消融实验

为了验证本文提出的子图检索模型的各个组件的有效性，在文物图像子图数据集上进行消融实验。测试内容主要包括：折叠多空洞金字塔池化有效性、注意力模块的有效性。

折叠多空洞金字塔池化有效性验证。本文使用ResNet50作为Backbone提取特征，使用折叠多空洞金字塔池化提取多尺度信息。为了验证折叠多空洞金字塔池化的有效性，本文分别使用折叠多空洞金字塔池化和普通多空洞金字塔池化进行测试，对比测试结果：基于普通多空洞金字塔池化模型的mAP为82.6%，基于折叠多空洞金字塔池化模型的mAP为85.3%，折叠多空洞金字塔池化比普通多空洞金字塔池化的平均检索精度mAP多了2.7%，证明了折叠多空洞金字塔池化的有效性。

注意力模块的有效性验证。本文使用注意力机制进行关键特征选择，避免无关特征对检索准确率的影响。为了验证该模块的有效性，本文将注意力模块去除，直接使用密集特征做相似度计算，对比测试结果，未引入注意力模块的子图检索模型的mAP为81.7%，引入注意力模块子图检索模型的mAP为85.3%，模型的检索有效性提高了3.6%。

（2）对比实验

图8所示为本文提出的子图检索模型在本文所构建的测试集上进行测试结果示意图，其中（a）和（c）为使用了原图像的部分图像进行检索，（b）和（d）为使用了某一文物的不同视角的完整图像进行检索。

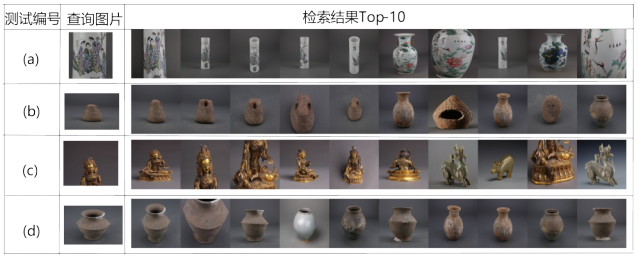

图8 子图检索算法检索结果Top-10

为了证明本文所提出的子图检索模型相比其他论文所提出方法在文物图像数据集上的有效性，本文选取了几个细粒度图像检索模型在构建的数据集上进行了实验。由于本文模型弱监督的特性，因此本文选取了几个弱监督细粒度图像检索模型进行对比。最终的检索效果对比如表1所示。

表1 不同模型检索精度对比表

检索模型	DELF[5]	DOLG[16]	DELG[9]	FMHPPA
mAP	80.9%	83.8%	81.6%	85.3%

5 结论

本文针对子图检索任务中因尺度变化和特征选择造成检索精度不高的问题，提出了一种融合折叠多空洞金字塔池化和注意力机制的文物图像的子图检索模型。该模型主要使用多空洞卷积和注意力机制进行特征提取和关键点选择，使用ASMK方法计算图像之间的相似度，最终实现文物图像子图检索的目标。为了验证算法的有效性，在所构建的子图数据集上进行了模型消融实验和性能对比试验，并与弱监督细粒度图像检索模型进行了对比，模型mAP达到了85.3%。由于文物图像受视角、姿态、光照变化等诸多原因影响造成特征有效表征的问题，未来将考虑引入多模态信息以提高模型检索精度。

参考文献（References）：

[1] 赵海英，周伟，侯小刚，等. 基于多任务学习的传统服饰图像双层标注[J]. 吉林大学学报(工学版)，2021，51(01)：293-302.

[2] 李明鑫，李雄飞，张金峰. 用于剪纸文化计算的数据存储模型[J]. 吉林大学学报(工学版)，2013，43(01)：152-157.

[3] 苗壮, 赵昕昕, 李阳, 等. 基于Swin Transformer的深度有监督哈希图像检索方法[J]. 湖南大学学报(自然科学版), 2023, 50(08): 62-71.

[4] Wei X S, Luo J H, Wu J, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881.

[5] Bai C, Li H, Zhang J, et al. Unsupervised adversarial instance-level image retrieval[J]. IEEE Transactions on Multimedia, 2021, 23: 2199-2207.

[6] Zheng X, Ji R, Sun X, et al. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 1226-1233.

[7] Noh H, Araujo A, Sim J, et al. Large-scale image retrieval with attentive deep local features[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 3456-3465.

[8] Phan L, Nguyen H T H, Warrier H, et al. Patch embedding as local features: unifying deep local and global features via vision transformer for image retrieval[C]// Proceedings of the Asian Conference on Computer Vision, 2022: 2527-2544.

[9] Cao B, Araujo A, Sim J. Unifying deep local and global features for image search[C]// Proceedings of Computer Vision-ECCV 2020: 16th European Conference, 2020: 726-743.

[10] Mohedano E, McGuinness K, O'Connor N E, et al. Bags of local convolutional features for scalable instance search[C]// Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016: 327-331.

[11] Salvador A, Giró-i-Nieto X, Marqués F, et al. Faster R-CNN features for instance search[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016: 9-16.

[12] Teichmann M, Araujo A, Zhu M, et al. Detect-to-retrieve: efficient regional aggregation for image search[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5109-5118.

[13] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, PMLR, 2011(15): 315-323.

[14] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[15] Zhao X, Pang Y, Zhang L, et al. Suppress and balance: a simple gated network for salient object detection[C]// Proceedings of Computer Vision-ECCV 2020: 16th European Conference, 2020: 35-51.

[16] Yang M, He D, Fan M, et al. Dolg: single-stage image retrieval with deep orthogonal fusion of local and global features[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 11772-11781.

编辑:赵志军