doi:10.3772/j.issn.1002-0470.2024.07.003

融合全局聚合与局部挖掘的建筑图像检索①

孟月波② 张紫琴 刘光辉 徐胜军

(西安建筑科技大学信息与控制工程学院 西安 710055)

摘 要 针对建筑图像易受到尺度变化和局部遮挡干扰而导致检索准确率低的问题,本文提出了一种融合全局聚合与局部挖掘的建筑图像检索网络。以 ResNet50 为骨干网络并在其后引入多尺度特征聚合的全局分支和注意力引导特征挖掘的局部分支,再通过正交融合策略高效整合双分支互补特征。其中,多尺度特征聚合模块结合混合空洞卷积和通道注意力对全局不同尺度的目标进行自适应加权聚合,增强网络对建筑多尺度显著特征的提取;注意力引导特征挖掘模块通过信息互补注意力对最显著特征标记擦除,实现对局部区域中潜在的细节信息的挖掘。所提方法在主流建筑数据集 ROxf 和 RPar 上的平均精度均值(mAP)指标分别达到了 81.54%(M)、62.43 %(H)和 90.28 %(M)、78.35%(H)。实验结果表明,该方法有效克服了尺度变化和局部遮挡的干扰,显著提升了建筑图像检索的准确率。

关键词 建筑图像; 图像检索; 特征聚合; 特征挖掘

建筑图像检索是计算机视觉领域中的研究热点之一,其主要目的是从建筑图片的数据库中检索出与查询图片语义信息相同的建筑图片,属于基于内容的图像检索(content based image retrieval, CBIR)[1]范畴。建筑检索在建筑物定位、地标识别、旅游导航等多个领域都具有重要的研究意义和广泛的应用前景。然而,建筑物经常受到局部遮挡、尺度变化、背景杂波等强干扰的影响,这些干扰给建筑图像检索的研究带来了诸多挑战[2]。

传统的图像检索包含特征提取和相似性度量 2 个子任务,首先通过手工设计的方法完成特征提取,然后利用度量学习算法约束特征间距离,更好地学习图片之间的相似性。提取强大的图像特征表示是图像检索的核心任务,早期常用的特征提取方法有尺度不变特征变换(scale invariant feature transform, SIFT)算法[3]、加速稳健特征(speeded up robust features, SURF)算法[4]、局部聚合描述子向量(vector of locally aggregated descriptors, VLAD)算法[5]等。这些方法需要耗费大量的人力成本,且提取的图像底层特征与高层语义信息之间存在“语义鸿沟”的问题,造成检索效率和精度普遍较低。随着深度学习的快速发展,图像检索的 2 个子任务被整合到一个统一的框架中,实现了端到端的图像检索过程,大幅提升了检索的性能[6-7]。根据深度学习网络输出特征的类型,建筑图像检索可以分为基于全局特征的检索方法、基于局部特征的检索方法和基于混合特征的检索方法。

全局特征的表征方式直观,擅长描述目标信息的整体形状和轮廓,因此基于深度全局特征图像检索的方法得到了研究人员的广泛关注。文献[8]提出了一种学习全局特征表示的实例级图像检索方法,通过聚合多个区域描述符为每个图像生成紧凑的全局描述符。文献[9]提出用于图像检索的注意力感知广义均值池化(attention-aware generalized

① 陕西省重点研发计划(2021SF-429)和陕西省自然科学基础研究计划(2023-JC-YB-532)资助项目。

② 女,1979 年生,博士,教授;研究方向:计算机视觉理解,建筑环境智能感知与调控,建筑智能化;联系人,E-mail: mengyuebo@163.com。

(收稿日期:2023-12-25)

mean pooling,AGeM)方法,将注意力机制与广义均值池化(generalized mean poding,GeM)结合生成注意力感知特征,通过归一化生成了紧凑的全局描述符,提升了建筑特征的表达能力。文献[10]提出了一种新的全局描述符 SOLAR(second-order loss and attention for image retrieval),利用空间注意力和特征描述符的相似性进行大规模图像检索,获得了图像全局特征的最佳描述。这些方法通过卷积神经网络获得了紧凑的全局描述符,有利于提升后续步骤中相似度的计算速度。然而,基于建筑整体信息的全局特征中包含背景等无关的杂乱信息,且对光照、遮挡、形变等干扰缺乏鲁棒性,因此,其检索效果有待进一步提高。

后续的研究发现局部特征可以保留更多的空间结构信息,更适用于图像检索任务。局部特征具有空间特征匹配度高的特点,能捕捉到区分不同建筑图像的关键信息。为了学习更多的细节特征,文献[11]提出了局部特征学习框架 DELF(deep local features),设计了一种用于大规模建筑图像检索的注意力局部特征描述符,实现了准确的特征匹配和几何验证。然而,DELF 缺少针对检索基准中感兴趣对象的边界框数据集,区域表示主要集中在统一区域选择或与类无关的区域选择上。因此,文献[12]提出了一种有效的图像检索区域聚集方法 D2R(detect-to-retrieve),利用训练好的地标检测器预测得到的边界框增强图像检索性能。这些方法保留建筑图片中更多具有判别性的细节信息,但是并不能充分描述目标物体的特征,检索效果仍然受到了限制。

单独学习全局特征或者局部特征都存在一定的局限性。最近的研究发现,与使用单一特征方法获得的结果相比,结合全局和局部特征的网络具有更高的检索准确性。因此,基于混合特征的检索方法成为了当前的主流方法。文献[13]提出 DELG(deep local and global features)模型,首先通过全局描述符搜索最相似的候选图片,然后利用局部特征的匹配结果进行重新排序。但是该模型采用的是两阶段检索方法,造成误差积累影响检索的性能。针对此问题,文献[14]提出 DOLG(deep orthogonal fusion of local and global features)模型,将局部特征和全局特征以单阶段方式相结合,有效避免了误差积累。注意力机制可以帮助模型关注有区分性的特征,有益于提升网络的检索性能。文献[15]提出 GLAM(global-local attention module)模型,结合了局部、全局、空间和通道注意力,从多个方面关注目标的显著特征,从而提高网络的检索精度。但是该模型使用多种注意力聚焦于图像中最具判别性的特征,网络结构过于复杂,并且缺乏对次显著局部细粒度特征的关注。文献[16]提出 DALG (deep attentive local and global modeling)模型,应用 Transformer 进行全局特征提取,设计基于窗口的多头注意力和空间注意力充分利用局部特征,并且引入可学习的交叉注意力模块跨层融合局部和全局特征,进行高效的单尺度图像检索。该模型通过可学习的融合策略自适应地利用两类特征间的互补性信息,进一步提高了图像检索的精度。然而,Transformer 的多头注意力机制内存占用过大,导致模型训练困难,并且仅使用单一尺度特征对图像进行检索,忽略了不同尺度特征对网络的贡献。

虽然混合特征在表达图像内容方面表现非常出色,但是它们在特征表达能力方面仍然受到限制,主要存在以下问题。(1)上述方法多数仅使用单一尺度特征对图像进行检索,忽略了全局上下文信息,导致特征提取不够全面,无法有效应对建筑目标的尺度变化的问题。(2)注意力机制虽然可以引导网络关注具有可区分性的部位,但是其通常只关注最显著的部分,而忽略了潜在的判别性细节信息,导致网络对局部遮挡等干扰的鲁棒性较差。鉴于此,本文基于 ResNet50 提出了一种融合全局聚合与局部挖掘的网络(global aggregation and local mining fusion network ,GALM-Net),旨在加强多尺度全局特征提取能力,充分挖掘潜在的多样化局部细节特征,从而改善最终输出特征图的质量,提高建筑图像检索的准确率。

1 相关理论

1.1 广义均值池化

广义均值池化(GeM)[17]是一种包含可学习参

数的池化方式，介于最大池化和平均池化之间。假设给定 GeM 池化一个特征图 $\boldsymbol{X}_P \in R^{C\times H\times W}$，其中 C、H 和 W 分别表示特征向量的通道数、高度和宽度。用 $\boldsymbol{X}_k \in R^{H\times W}$ 表示特征图 $\boldsymbol{X}_P$ 通道维度上的第 k 张特征图，其中 $k \in \{1\cdots C\}$，向量 $\boldsymbol{f}$ 作为该池化操作的输出，计算公式如下：

$$\boldsymbol{f} = [\boldsymbol{f}_1\cdots\boldsymbol{f}_k\cdots\boldsymbol{f}_C]^{\mathrm{T}},\ \boldsymbol{f}_k = \left(\frac{1}{|\boldsymbol{X}_k|}\sum_{\boldsymbol{x}\in\boldsymbol{X}_k}\boldsymbol{x}^{p_k}\right)^{\frac{1}{p_k}} \quad (1)$$

由式(1)可以看出，当 $p_k \to \infty$ 时，GeM 近似于最大池化操作；当 $p_k \to 1$ 时，GeM 近似于平均池化操作。GeM 池化兼顾了最大池化和平均池化的优点，能够更好地挖掘特征之间的判别信息。更重要的是，GeM 池化是一种可微操作，可以通过端到端的方式训练整个网络求取最佳 p_k。

1.2　正交融合策略

正交融合策略[14]可以对互补的全局特征与局部特征进行精细准确的高效融合，使每个局部特征点排除与全局特征相关的信息，在降低深层特征之间相关性的同时丰富特征的多样性，使建筑特征的表达更加简洁和全面。

假设网络提取到的全局特征为 $\boldsymbol{f}_g \in R^{C\times H\times W}$，局部特征为 $\boldsymbol{f}_l \in R^{C\times H\times W}$。在正交融合模块中，将局部特征 $\boldsymbol{f}_l^{(x,y)}$ 分解为 2 个向量，一个与全局特征平行 $\boldsymbol{f}_{l,\mathrm{proj}}^{(x,y)}$，另一个与全局特征正交 $\boldsymbol{f}_{l,\mathrm{orth}}^{(x,y)}$。经过高维正交分解后，每个局部特征点在全局特征上的投影可以表示为

$$\boldsymbol{f}_{l,\mathrm{proj}}^{(x,y)} = \frac{\boldsymbol{f}_g\cdot\boldsymbol{f}_l^{(x,y)}}{|\boldsymbol{f}_g|^2}\boldsymbol{f}_g \quad (2)$$

式中，$|\boldsymbol{f}_g|^2$ 是全局特征的 L_2 范式，$\boldsymbol{f}_g\cdot\boldsymbol{f}_l^{(x,y)}$ 是点积运算。

局部特征 $\boldsymbol{f}_l^{(x,y)}$ 通过计算与其在全局特征上的投影分量 $\boldsymbol{f}_{l,\mathrm{proj}}^{(x,y)}$ 之间的差值，得到局部特征的正交分量 $\boldsymbol{f}_{l,\mathrm{orth}}^{(x,y)}$ 的表达式为

$$\boldsymbol{f}_{l,\mathrm{orth}}^{(x,y)} = \boldsymbol{f}_l^{(x,y)} - \boldsymbol{f}_{l,\mathrm{proj}}^{(x,y)} \quad (3)$$

正交融合模块整体流程如图 1 所示，将全局特征 $\boldsymbol{f}_g$ 和局部特征 $\boldsymbol{f}_l$ 作为模块的输入，从局部特征中提取与全局特征正交的分量。然后将局部特征的正交分量与全局特征进行融合，生成包含丰富信息的最终特征向量 $\boldsymbol{f}$。

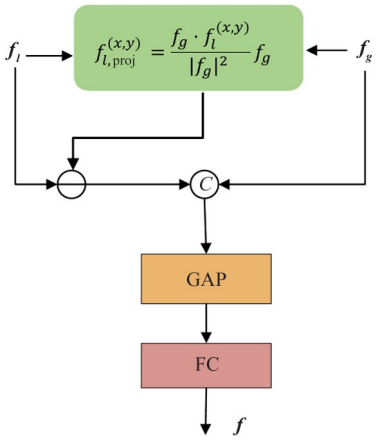

图 1　正交融合模块

2　融合全局聚合与局部挖掘的建筑图像检索网络

2.1　网络整体概述

融合全局聚合与局部挖掘的建筑图像检索网络 GALM-Net 的网络框架结构如图 2 所示，主要包括特征提取骨干网络 ResNet50、多尺度特征聚合(multi-scale feature aggregation，MSFA)的全局分支、注意力引导特征挖掘(attention-guided feature mining，AGFM)的局部分支和正交融合模块(orthogonal fusion module，OFM)4 个部分。

ResNet50[18]是图像检索常用主干网络之一，通过引入残差学习有效地解决了深度网络的退化问题，且提取的深层特征蕴含丰富的语义信息。因此，本文选择 ResNet50 作为特征提取的主干网络，通过在 4 个残差层后引入全局分支和局部分支，捕获建筑图像丰富且全面的特征。在全局分支中，设计了一个多尺度的特征聚合模块，通过混合空洞卷积增大感受野以适应多尺度目标，在通道注意力机制的指导下对全局不同尺度的目标进行动态聚合，获得包含丰富上下文信息的多尺度显著特征。然后，利用 GeM 池化和全连接层(full connection，FC)对通道数进行降维，得到全局分支网络的特征表示 $\boldsymbol{f}_g$。在局部分支中，提出了一种注意力引导的特征挖掘模块，通过信息互补注意力对最显著特征标记，并根据阈值选择操作进行特征擦除，迫使网络充分挖掘局部区域中多个潜在且具有辨别性的细节信息，有效提高建筑特征在遮挡条件下的表达能力。然后，利

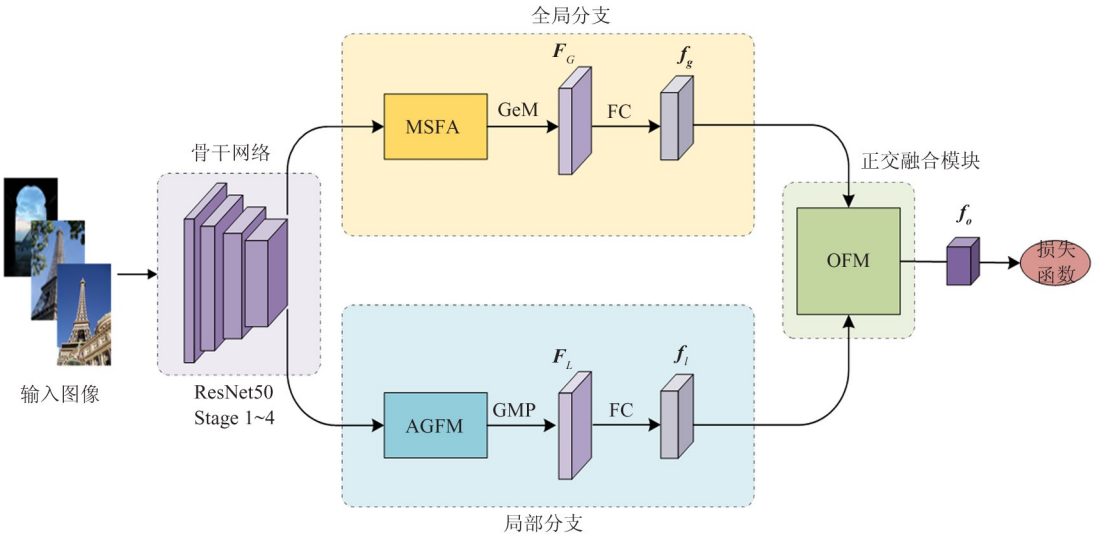

图 2　网络基本结构

用全局最大池化(global max pooling,GMP)和 FC 层对通道数进行降维,得到局部分支网络的特征表示 f_l。接着,通过正交融合策略对全局分支和局部分支提取到的互补特征进行高效整合,消除信息间的冗余,得到了网络最后的输出特征 f_o。最后,使用联合损失函数对网络进行监督训练,提升建筑图像检索的准确率。

2.2　多尺度特征聚合模块

由于建筑图像的尺度变化较大,图像检索存在全局上下文信息提取困难、建筑特征较难提取等问题。基于此,本文设计了多尺度特征聚合模块,由多尺度特征提取和通道聚合 2 部分构成,其结构如图 3所示。多尺度特征提取部分通过模拟人类视觉系统中人眼感受野与离心率变化之间的关系[19],加强网络对全局范围内重要特征的提取能力;通道聚合部分对不同尺度的特征进行加权关注,使得网络能够根据输入图片自适应地调整感受野的大小,从而得到更加全面且充分的特征信息。

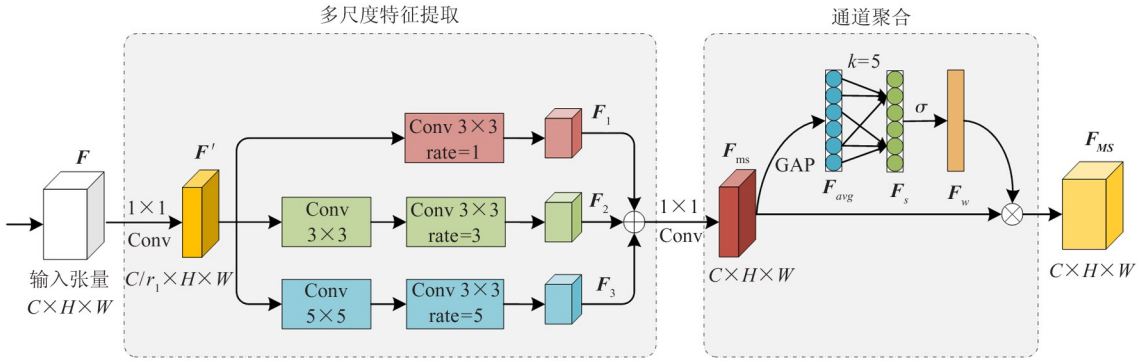

图 3　多尺度特征聚合模块

2.2.1　多尺度特征提取

人类的视觉系统是由多个具有不同感受野的部分复合而成的,靠近中心位置的特征通常更重要。然而一般的卷积网络通常采用固定大小的感受野提取特征,失去了对不同视野的分辨能力,导致模型对输入图片中目标尺度变化的适应性较差。针对此问题,本文引入空洞卷积(dilated convolutions)[20]扩大感受野,采用多尺度特征提取的方法捕捉不同尺度特征,获得更广泛的上下文信息。具体实现过程如下。

首先,使用卷积核大小为 1×1 的卷积对输入特征图 $\boldsymbol{F}\in R^{C\times H\times W}$ 在通道维度进行降维操作以减少

参数量,降维后的特征图为 $\boldsymbol{F}' \in R^{C/r_1\times H\times W}$。然后,利用 3 组不同空洞率的空洞卷积有效扩大特征图的感受野,分别得到 3 组不同尺度的特征 $\boldsymbol{F}_1 \in R^{C/r_1\times H\times W}$、$\boldsymbol{F}_2 \in R^{C/r_1\times H\times W}$、$\boldsymbol{F}_3 \in R^{C/r_1\times H\times W}$。通过式(4)将它们按元素相加后再使用 1×1 卷积调整通道数,获得多尺度特征图 $\boldsymbol{F}_{ms} \in R^{C\times H\times W}$。

$$\boldsymbol{F}_0 = \boldsymbol{F}_1 \oplus \boldsymbol{F}_2 \oplus \boldsymbol{F}_3 \tag{4}$$

式中,⊕表示逐元素相加操作。

2.2.2　通道聚合

建筑图像具有拍摄视角多变的特点,同类建筑中图像的尺度信息分布不均,每个尺度的重要程度也不同。尽管对建筑图像进行多尺度特征提取获得了丰富的上下文信息,但是采用简单的通道拼接(concat)或相加(add)使得网络平等地对待每个尺度的特征,不利于增强多尺度特征中关键建筑元素特征信息的有效表达。因此,本文采用有效通道注意力网络 ECANet(efficient channel attention network)[21]以更好地获取特征通道间的相关性,促进网络自适应引导多尺度信息的动态聚合,减少信息冗余并提高重要特征表达的能力。

ECANet 通过快速一维卷积实现了无降维的局部跨通道交互,在保证性能的同时显著降低了模型复杂度。具体地,将多尺度特征图 $\boldsymbol{F}_{ms} \in R^{C\times H\times W}$ 进行全局平均池化得到聚合特征 $\boldsymbol{F}_{\text{avg}} \in R^{C\times 1\times 1}$,然后使用卷积核大小为 k 的一维卷积实现局部跨通道信息交互得到特征图 $\boldsymbol{F}_s \in R^{C\times 1\times 1}$,可表示为

$$\boldsymbol{F}_s = \boldsymbol{f}_{1D}^{k\times k}(\boldsymbol{F}_{\text{avg}}) \tag{5}$$

式中,$\boldsymbol{f}_{1D}^{k\times k}$ 表示卷积核大小为 k 的一维卷积,k 代表局部跨通道相关性学习的交互范围,由通道数 C 自适应确定,k 与 C 之间存在一个映射关系,可表示为

$$k = \psi(C) = \left|\frac{\log_2 C}{2} + \frac{1}{2}\right|_{\text{odd}} \tag{6}$$

式中,$|\,|_{\text{odd}}$ 表示离 k 最近的奇数。

然后将特征图 $\boldsymbol{F}_s$ 通过 Sigmoid 激活函数归一化得到通道权重矩阵 $\boldsymbol{F}_w \in R^{C\times 1\times 1}$,可表示为

$$\boldsymbol{F}_w = \sigma(\boldsymbol{F}_s) \tag{7}$$

式中,σ 表示 Sigmoid 激活函数。

最后将生成的通道权重矩阵 $\boldsymbol{F}_w$ 与多尺度特征图 $\boldsymbol{F}_{ms}$ 逐元素相乘得到 $\boldsymbol{F}_{MS} \in R^{C\times H\times W}$,可表示为

$$\boldsymbol{F}_{MS} = \boldsymbol{F}_w \otimes \boldsymbol{F}_{ms} \tag{8}$$

式中,⊗表示逐元素相乘。

2.3　注意力引导的特征挖掘模块

当图像中的目标建筑被其他建筑、树木或行人遮挡时,一些具有描述性的建筑特征被抑制,通过上述多尺度特征聚合模块得到的全局显著特征所提供的信息失去了可判别性。相比于全局特征,局部特征可以从建筑物屋顶或窗户等区域获取辨别性信息,具有良好的抗干扰能力。因此,本文提出了信息互补注意力(information complements attention,ICA)来学习最具鉴别力的局部特征表示,并减轻来自背景的干扰。然而在局部特征的学习过程中,网络通常只关注最显著的部分,从而忽略了其他潜在的细节特征,而这些特征在一些具有挑战性的场景中是重要的线索。若仅仅依靠最具判别力的特征进行图像检索,容易使模型对目标建筑的屋顶、窗户、遮挡物等显著特征过拟合。为了缓解上述问题,本文提出了注意力引导的特征挖掘模块,充分挖掘易被忽视的细粒度局部特征,增强模型对多样性特征的学习能力,提升模型的泛化性和鲁棒性。

注意力引导的特征挖掘模块结构如图 4 所示。首先,该模块将从骨干网络中提取的特征图 $\boldsymbol{F} \in R^{C\times H\times W}$ 作为输入,然后经过信息互补注意力提取与建筑相关的感兴趣区域注意力图 $\boldsymbol{F}_A \in R^{C\times H\times W}$,帮助网络学习包含丰富信息的细粒度局部特征。接着通过阈值选择操作(threshold)引导网络擦除特征图中最具鉴别性的部分,强迫网络挖掘剩余区域中的潜在特征并获得相应的擦除掩码 $\boldsymbol{F}_D \in R^{C\times H\times W}$。最后将擦除掩码与输入特征图进行逐元素相乘以生成擦除特征图 $\boldsymbol{F}_{AD} \in R^{C\times H\times W}$,通过擦除显著特征的方式鼓励神经网络更多地关注次显著细节特征。

2.3.1　信息互补注意力

为了有效捕获空间维度和通道维度上的重要特征信息,本文设计了信息互补注意力模块 ICA。信息互补注意力由并行的空间注意力分支(spatial attention,SA)和通道注意力分支(channel attention,CA)组成,通道注意力有助于模型确定输入特征图中哪些内容具有重要的区分作用,空间注意力更加

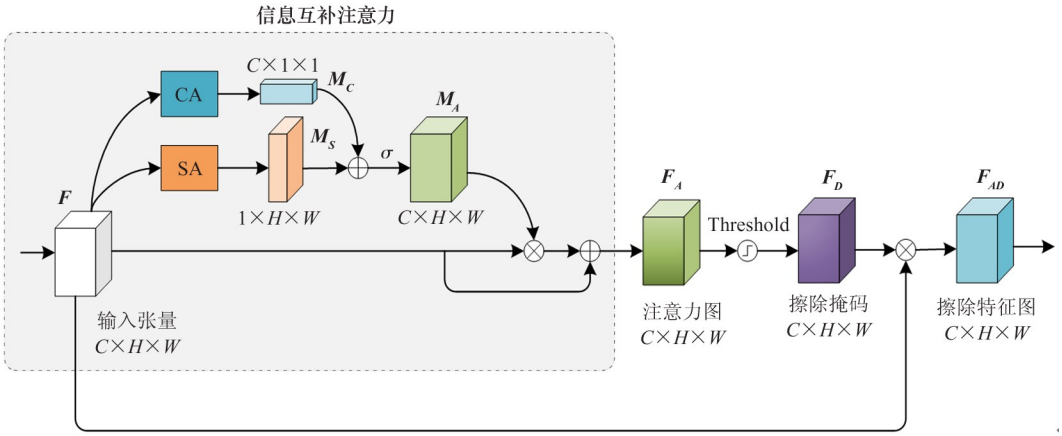

图 4　注意力引导的特征挖掘模块

关注图像中关键信息的空间位置。将通道和空间分支相结合形成不同维度上的信息互补,捕捉建筑物的空间特征以及重要信息,从而更准确地区分目标建筑和背景。ICA 通过对特征图的权重进行调整,自适应地选择主要的通道和空间区域,在保留有用特征的同时抑制无关信息的干扰,可以进一步提高模型的特征信息的表征能力和泛化性能。具体实现步骤如下:

输入特征图 $\boldsymbol{F} \in R^{C\times H\times W}$ 分别经过 2 种注意力模块后得到通道注意力特征图 $\boldsymbol{M}_C \in \boldsymbol{R}^{C\times H\times W}$ 和空间注意力特征图 $\boldsymbol{M}_S \in \boldsymbol{R}^{C\times H\times W}$,通过逐元素相加操作得到融合的注意力权重向量 $\boldsymbol{M}_A \in R^{C\times H\times W}$,最后将与输入特征进行逐元素相乘,并添加残差连接再次与输入特征相加,得到特征重构后的加权注意力特征图 $\boldsymbol{F}_A \in R^{C\times H\times W}$。

$$\boldsymbol{M}_A = \sigma(\boldsymbol{M}_C + \boldsymbol{M}_S) \tag{9}$$

$$\boldsymbol{F}_A = \boldsymbol{M}_A \otimes \boldsymbol{F} + \boldsymbol{F} \tag{10}$$

式中,σ 表示 Sigmoid 函数,⊗表示逐元素相乘。

(1) 通道注意力

通道注意力分支利用通道之间的相互依赖性进行建模,增强各种建筑特征的表征能力,其结构如图 5所示。全局平均池化(global average pooling, GAP)和全局最大池化(global max pooling, GMP)是深度学习中的 2 个基本操作,GAP 可以很好地保留结构信息,但很容易被背景杂波干扰,而 GMP 以失去结构信息为代价,通过关注最突出的部分克服了这个问题。为了更好地聚合判别信息,本文的通道注意力分支采用双通道池化层的方式联合使用GAP 和 GMP。通道注意力机制的具体实现步骤如下。

首先,输入特征图 $\boldsymbol{F} \in R^{C\times H\times W}$ 分别通过 GAP 和 GMP 获得 2 个全局信息描述符 $\boldsymbol{M}_{\text{avg}} \in R^{C\times 1\times 1}$ 和 $\boldsymbol{M}_{\max} \in R^{C\times 1\times 1}$。然后,通过 2 个全连接层获得注意力权重矩阵 $\boldsymbol{A}_C \in R^{C\times 1\times 1}$,用于学习输入图像中不同通道之间的重要性。最后,经过 Sigmoid 激活函数,获得了输入特征层每一个通道的权值(0 ~ 1 之间),生成最终的通道注意力特征图 $\boldsymbol{M}_C \in R^{C\times 1\times 1}$。

$$\boldsymbol{M}_C = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{M}_{\text{avg}}) + (\boldsymbol{W}_2\delta\boldsymbol{M}_{\max})) \tag{11}$$

式中,σ 表示 Sigmoid 函数,δ 表示 ReLU 函数,$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别表示 2 个全连接层的权重矩阵,⊗表示逐元素相乘。

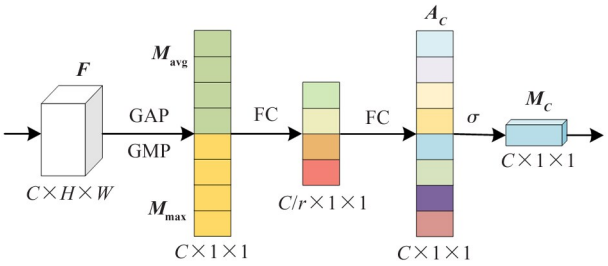

图 5　通道注意力

(2) 空间注意力

空间注意力分支通过对输入特征图中不同位置的信息进行加权融合,帮助网络确定关键信息在输入图像中的位置分布,其结构如图 6 所示。考虑到上下文信息对空间位置的理解至关重要,因此在空间注意力分支中使用空洞卷积来扩大感受野,构建更有效的空间图,提高网络对不同区域之间位置关系的理解和分析能力。空间注意力机制的具体实现

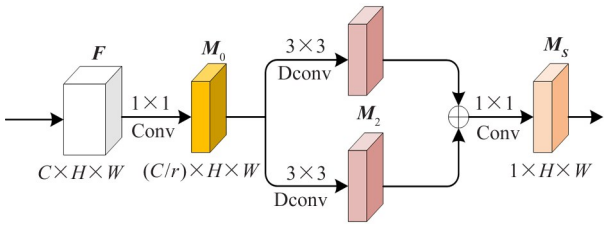

图 6　空间注意力

步骤如下。

首先,对输入特征图进行卷积核为 1 ×1 的卷积操作,得到通道降维后的特征 $\boldsymbol{M}_0 \in R^{(C/r)\times H\times W}$。然后,应用 2 个并联的卷积核大小为 3 ×3、空洞率为 4 的空洞卷积,得到 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 并将结果相加,用于聚合具有较大感受野的上下文信息。最后,使用卷积核为 1 ×1 的卷积获得空间注意力图 $\boldsymbol{M}_S \in R^{1\times H\times W}$。

$$\boldsymbol{M}_S = f_3^{1\times1}(f_2^{3\times3}f_0^{1\times1}(\boldsymbol{F}) + f_1^{3\times3}f_0^{1\times1}(\boldsymbol{F})) \quad (12)$$

式中,f 表示卷积运算,f 的上标表示卷积核的大小。

2.3.2　阈值选择操作

通过信息互补注意力获得了注意力特征图 $\boldsymbol{F}_A \in R^{C\times H\times W}$,假设 $\boldsymbol{F}_A(x, y)$ 是注意力特征图 $\boldsymbol{F}_A$ 中的任意一个像素点,$\boldsymbol{F}_A$ 中最大像素值设为 $\boldsymbol{F}_{\max}$,并设置阈值参数 θ 的取值范围为(0,1),则阈值为 $T = \theta \times \boldsymbol{F}_{\max}$。将注意力特征图中的每一个点都与阈值 T 进行比较,如果注意力特征图中某一个像素点的值大于阈值 T,则擦除掩码 $\boldsymbol{F}_D \in R^{C\times H\times W}$ 中对应点的值设置为 0,否则,将其设置为 1。擦除掩码中值为 0 的点组成的区域即为遮挡区域,公式如下:

$$\boldsymbol{F}_D(x, y) = \begin{cases} 0 & \boldsymbol{F}_A(x, y) > T \\ 1 & \text{其他} \end{cases} \quad (13)$$

式中,$x \in [0,H)$,$y \in [0,W)$,T 表示阈值

2.4　联合损失函数

本文的研究旨在解决图像检索任务,该任务既包含分类任务又包含度量学习。为了使模型在训练过程中获得更好的表征学习能力,本文结合交叉熵损失 L_s 和三元组损失 L_t 作为联合损失函数 L 来优化网络。联合损失函数的表达式为

$$L = L_s + L_t \quad (14)$$

交叉熵损失函数被广泛运用于图像分类的任务中,通过最小化真实概率分布与预测概率分布之间的差异对网络进行优化。在建筑图像检索中,每张建筑图片对应一个类别标签,因此可以转化为一个分类问题。交叉熵损失表达式为

$$L_s = -\sum_{i=1}^{N} \log \frac{e^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i}}{\sum_{k=1}^{C} e^{\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{f}_i}} \quad (15)$$

式中,N 表示一个批次中的图像数量,$\boldsymbol{f}_i$ 表示第 i 个学到的特征,$\boldsymbol{W}_k$ 表示对应类别的权重向量,C 表示训练集中建筑的类别数。

三元组损失函数用来度量 2 张图像之间的相似性。一个三元组 $(\boldsymbol{X}_a, \boldsymbol{X}_p, \boldsymbol{X}_n)$ 由锚点样本 $\boldsymbol{X}_a$、正样本 $\boldsymbol{X}_p$、负样本 $\boldsymbol{X}_n$ 构成,目的是使正样本对的距离小于负样本对的距离,增大类间差异并减小类内差异。三元组损失函数表达式为

$$L_t = \max(0, m + d(r_a, r_p) - d(r_a, r_n)) \quad (16)$$

式中,$d(r_a, r_p)$ 代表锚点样本与正样本之间的距离,$d(r_a, r_n)$ 代表锚点样本与负样本之间的距离,m 表示正负样本距离的差值边界 margin。

3　实验数据与分析

3.1　数据集与评价指标

实验数据集采用公共数据集 ROxf[22] 和 RPar[22]。ROxf 数据集中包含 4 933 张 1 024 ×768 像素大小的图片,由牛津万灵学院、牛津基督教堂、牛津大学等 11 类牛津地标建筑组成。RPar 数据集中包含 6 322 张 1 024 ×768 像素大小的图片,由埃菲尔铁塔、凯旋门、卢浮宫、巴黎圣母院等 12 类巴黎地标建筑组成。ROxf 数据集和 RPar 数据集是基于 Oxford5K[23] 数据集和 Paris6K[24] 数据集重新修订的版本,为图像分配了新的 4 类标签:Easy、Hard、Unclear 和 Negative。在此基础上,使用不同的标签组合作为正面图像来设置检索难度级别,分别是“Easy”、“Medium”和“Hard”。其中,Easy(E)级别是将标记为 Easy 图片设置为正面图像;Medium(M)级别是将标记为 Easy 和 Hard 的图片设置为正面图像;Hard(H)级别是将标记为 Hard 的图片设置为正面图像。由于 E 级别的难度设置接近原始数据集,导致图像检索的挑战性较小,因此本文的实验部分均采用 M 级别和 H 级别进行实验评估。

为了与最新的方法进行公平的比较,本文采用

平均精度均值(mean average precision,mAP)作为网络性能的评价指标。mAP 的取值范围是[0,1],mAP 越大则表明网络的检索精度越高。

3.2 实验细节

本文实验采用 PyTorch-1.7 深度学习框架,GPU 型号为 NVIDIA GeForce RTX 2080Ti,环境配置为 CUDA11.0 + cuDNN8.0 + Python3.7.4。在模型的训练过程中,通过随机裁剪和随机水平翻转进行数据增强。训练批次设置为 32,总共训练 200 轮,采用随机梯度下降法进行优化更新,动量设置为 0.9,权重衰减系数设置为 0.000 1。训练过程中的损失下降曲线图如图 7 所示,横轴表示训练的迭代次数,纵轴表示损失值的大小。可以看出,网络在训练前期损失值下降较快,当训练迭代次数达到 100 左右时损失值下降曲线趋于稳定,最终收敛在 0.2 附近,说明本文设计的模型训练结果较为理想。对于 1.1 节中的广义均值池化,将参数 p_k 固定为 3;对于 2.2.2节中的通道注意力 ECANet,将局部跨通道相关性学习的交互范围 k 设置为5;对于2.4 小节中的三元组损失,将正负样本距离的差值边界 m 设置为 0.3;对于 2.3.3 小节中的阈值选择操作,将阈值参数 θ 设置为 0.8。

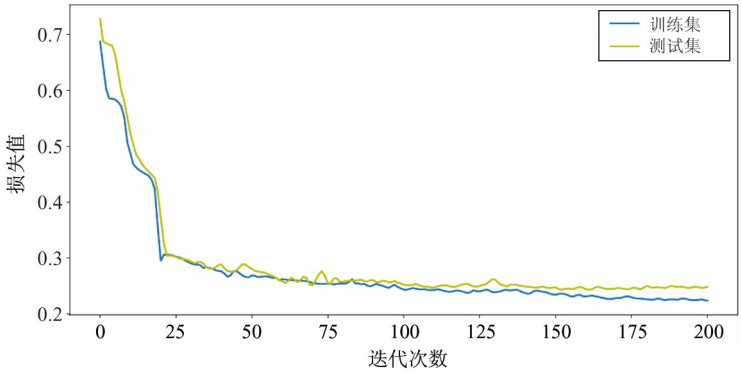

图 7 网络训练过程中损失下降曲线图

3.3 对比实验与分析

为了验证本文所提方法的有效性,表 1 给出了 GALM-Net 在 ROxf 数据集和 RPar 数据集上与其他先进的建筑图像检索算法的对比结果。为了与先进算法进行公平的比较,对比算法均采用 ResNet50 作为主干网络,并且除了本文特有的超参数外,其他实验参数设置均与本文算法保持一致,具体参见 3.2 节。对比的主流方法包括 3 种类型:基于全局特征的方法(AGeM、SOLAR)、基于局部特征的方法(D2R、HOW)、基于混合特征的方法(DELG、DOLG、

表 1 不同算法结果对比实验(%)

类型	方法	ROxf		RPar	
		M	H	M	H
全局特征	AGeM[9]	67.00	40.70	78.10	57.30
	SOLAR[10]	69.90	47.90	81.60	64.50
局部特征	D2R[12]	76.00	52.40	80.20	58.60
	HOW[6]	79.40	56.90	81.60	62.40
全局特征和局部特征	DELG[13]	75.40	53.70	82.20	63.70
	DOLG[14]	80.50	58.82	89.81	77.70
	GLAM[15]	78.60	60.20	88.50	76.80
	DALG[16]	78.01	54.40	88.97	76.35
	CVNet[7]	81.00	62.10	88.80	76.50
	GALM-Net	81.54	62.43	90.28	78.35

GLAM、DALG、CVNet)。

实验结果表明,相较于仅使用单一特征的方法,本文方法的实验结果在 2 个公开数据集上均有显著提升,表现出良好的性能。进一步与融合全局和局部特征的方法相比,本文方法在 2 个数据集上均取得较高的精度。具体来说,本文在 ROxf 数据集上 mAP 指标达到了 81.54%(M)和 62.43%(H),在 RPar 数据集上的 mAP 指标达到了 90.28 %(M)和 78.35%(H)。其中与当前最先进的算法 DOLG 和 CVNet 相比,在数据集 ROxf(M)、ROxf(H)、RPar(M)、RPar(H)上的 mAP 分别提高了 0.54、0.33、0.47、0.65 个百分点。可以看出,本文所提方法的性能超过了当前大多数主流算法并且在总体上优于最先进的算法。

3.4 消融实验及分析

3.4.1 不同模块的性能分析

为了探究不同模块对建筑图像检索性能的影响,本文分别在 ROxf 数据集和 RPar 数据集上进行消融实验,结果如表 2 所示。其中,w/o MSFA 表示去掉多尺度特征聚合模块,w/o AGFM 表示去掉注意力引导的特征挖掘模块,w/o OFM 表示去掉正交融合模块,Full model 表示同时使用上述 3 个模块。

表 2 不同模块的消融实验(%)

模块	ROxf		RPar	
	M	H	M	H
w/o MSFA	79.67	60.94	89.13	76.70
w/o AGFM	80.16	61.82	89.17	77.26
w/o OFM	80.65	62.18	89.75	78.02
Full model	81.54	62.43	90.28	78.35

由消融实验结果可知,3 个模块对建筑检索的性能提升都有不同程度的贡献,具体分析如下。

(1)去掉多尺度特征聚合模块,ROxf 数据集和 RPar 数据集的 mAP 分别为 79.67%(M)、60.94%(H)和 89.13%(M)、76.70%(H),添加多尺度特征聚合模块之后,2 个数据集上 mAP 分别提高了 1.87(M)、1.49(H)和 1.15(M)、1.65(H)个百分点。这表明 MSFA 模块可以有效获取并整合在不同感受野下提取到的多尺度特征,缓解了建筑尺度变化对检索性能带来的负面影响。

(2)去掉注意力引导的特征挖掘模块,ROxf 数据集和 RPar 数据集的 mAP 分别为 80.16%(M)、61.82%(H)和 89.17%(M)、77.26%(H),添加注意力引导的特征挖掘模块之后,2 个数据集上 mAP 分别提高了 1.38(M)、0.61(H)和 1.11(M)、1.09(H)个百分点。这表明 AGFM 模块可以学到目标建筑多个潜在且关键的局部特征,增强了建筑特征在局部遮挡条件下的鲁棒性。

(3)去掉正交融合模块,ROxf 数据集和 RPar 数据集的 mAP 分别为 80.65%(M)、62.18%(H)和 89.75%(M)、78.02%(H),添加正交融合模块之后,2 个数据集上 mAP 分别提高了 0.89(M)、0.25(H)和 0.53(M)、0.33(H)个百分点。这表明 OFM 模块可以有效融合不同类型的特征,网络提取到的特征也更加丰富。当 3 个模块同时使用时,相较于仅使用单一模块进行图像检索,数据集 RPar 和数据集 ROxf 的性能均达到最佳,进一步提升了建筑图像检索的准确率。

3.4.2 不同注意力的性能分析

为了验证本文所提的信息互补注意力模块的有效性,将局部分支中添加的 ICA 模块使用不同类型的注意力模块进行替换,分别在 ROxf 数据集和 RPar 数据集上进行消融实验,实验结果如表 3 所示。其中,SE 是通道注意力,ECA(efficient channel attention)是对 SE 改进的通道注意力,CBAM(convolutional block attention module)是通道与空间混合注意力。

表 3 不同类型注意力的消融实验(%)

注意力	ROxf		RPar	
	M	H	M	H
SE[25]	80.25	61.73	89.07	76.49
ECA[21]	80.76	61.59	89.15	76.84
CBAM[26]	81.30	62.06	89.81	77.46
ICA	81.54	62.43	90.28	78.35

从表 3 的消融实验结果可知,加入 ICA 模块的网络在 2 个数据集上的 mAP 达到最佳。在 ROxf 数据集上,使用 CBAM 时的 mAP 为 81.30%(M)、

62.06%(H),而使用 ICA 的 mAP 为 81.54%(M)、62.43%(H),分别提升了 0.24(M)、0.37(H)个百分点;在 RPar 数据集上,使用 CBAM 时的 mAP 为 89.81%(M)、77.46%(H),而使用 ICA 的 mAP 为 90.28%(M)、78.35%(H),提升了 0.47(M)、0.89(H)个百分点。ICA 能够将通道和空间维度捕获的信息进行互补,保留有用信息的同时抑制无关特征,因此在网络中加入 ICA 更有助于提取全面的显著特征。

3.4.3 混合空洞率的选择

多尺度特征聚合模块中空洞率的选择对特征的提取效果存在影响,若空洞率过大则会在卷积的操作过程中引入噪声并破坏特征的连续性,容易引起网格效应;若空洞率过小则会影响网络对特征信息的提取能力。因此,根据建筑图像检索任务的特点和混合空洞卷积的设计原则,使用 3 组不同的空洞率组合进行消融实验,实验结果如表 4 所示。

表 4 不同空洞率对性能的影响(%)

空洞率	ROxf		RPar	
	M	H	M	H
(1,2,3)	80.37	61.76	89.52	77.02
(1,3,5)	81.54	62.43	90.28	78.35
(1,5,7)	80.20	61.25	89.83	77.64

实验结果表明,使用(1,3,5)组合的混合空洞率时,网络性能最好。与(1,2,3)组合相比,ROxf 数据集和 RPar 数据集的 mAP 分别提高了 1.17(M)、0.67(H)和 0.76(M)、1.33(H)个百分点。与(1,5,7)组合相比,ROxf 数据集和 RPar 数据集的 mAP 分别提高了 1.34(M)、1.18(H)和 0.45(M)、0.71(H)个百分点。最佳的组合方式为(1,3,5),这样设计可以保证感受野能够覆盖特征图中的每个像素点,有效地收集不同区域的特征信息。

3.4.4 阈值参数的选择

注意力引导的特征挖掘模块通过设置不同的擦除阈值来模拟建筑物被遮挡的真实场景,其中阈值参数 $\theta \in (0,1)$,若 θ 取 0,则阈值 T 为 0,即擦除特征图的全部特征;若 θ 取 1,则阈值 T 为特征图中的最大像素值 F_{max},即没有擦除显著特征。θ 具体的取值因任务而异,为了找到最合适建筑图像检索任务的阈值大小,在 ROxf 数据集和 RPar 数据集上分别设置不同的阈值进行消融实验,实验结果如表 5 所示。

表 5 不同 θ 对性能的影响(%)

θ	ROxf		RPar	
	M	H	M	H
0.9	81.35	62.37	90.12	78.24
0.8	81.54	62.43	90.28	78.35
0.7	81.02	62.13	90.06	78.15
0.6	80.86	61.75	89.86	77.91
0.5	80.70	61.44	89.50	77.63
0.4	80.55	61.23	89.40	77.56
0.3	80.30	61.18	89.36	77.48
0.2	80.27	61.12	89.34	77.33
0.1	80.21	61.09	89.22	77.25

实验结果表明,当 θ 取 0.8 时,网络实现了最佳性能。与 θ 取 0.9 时相比,ROxf 数据集和 RPar 数据集的 mAP 分别提高了 0.19(M)、0.06(H)和 0.16(M)、0.11(H)个百分点。与 θ 取 0.1 时相比,ROxf 数据集和 RPar 数据集的 mAP 分别提高了 1.33(M)、1.34(H)和 1.06(M)、1.10(H)个百分点。若阈值设置过高则擦除区域太小,会保留过多的无关特征;设置过低则擦除区域太多,会丢失目标建筑信息。因此,针对建筑图像检索任务,阈值参数 θ 的大小取 0.8 最为合适。

3.5 可视化分析

3.5.1 特征图可视化

为了直观地展示每个模块的作用,本文从数据集 ROxf 和 RPar 中各选取一张图片作为示例样本进行可视化实验。对主干网络 ResNet50,以及引入 MSFA 和 AGFM 后的特征图进行可视化分析,结果如图 8 所示。

可以看出,对于主干网络 ResNet50,模型只能识别到建筑图像中最具判别性的显著特征,如建筑物的屋顶。当引入 MSFA 后,不同尺度语义信息进行融合增强了模型关注区域的丰富性,但是仍然忽略了一些关键的细节信息。当添加 AGFM 之后,模型学习到多个潜在的局部特征,在保证模型丰富性的同时挖掘到更多具有辨别性的次显著特征。当

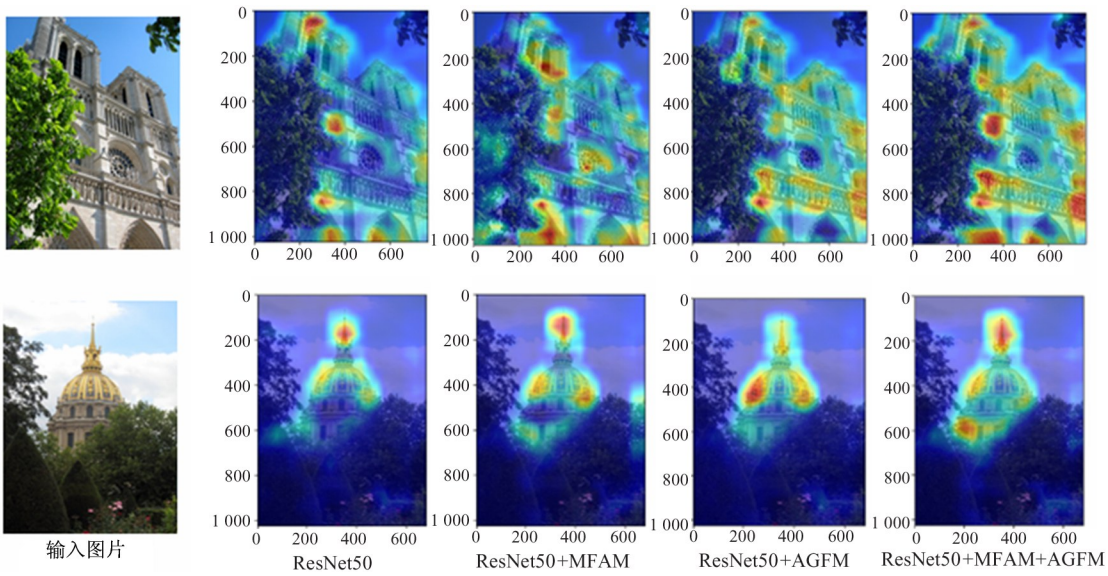

图 8　特征图可视化

MFAM 和 AGFM 同时使用时，可以看出特征图可视化中建筑的全局轮廓更清晰，并且准确覆盖到更多的细节特征。可视化结果表明，建筑特征的提取更加多样且完整，验证了本文所提方法的有效性。

3.5.2　检索结果可视化

在 ROxf 数据集和 RPar 数据集中，分别选取具有局部遮挡和尺度变化的样本作为检索结果可视化的示例图片。查询结果如图 9 所示，共包含 11 列图像，其中第 1 列图像是查询图像，后 10 列图像显示了网络查询后返回的前 10 个匹配结果。使用实线边框标记检索正确的图片，使用虚线边框标记检索错误的图片。查询结果表明，相比于 ResNet50 网络，本文所提方法取得了更高的检索准确率。在一些建筑尺度变化较大和局部遮挡的情况下，GALM-Net 仍然能较好地识别出特定类别的建筑。

4　结　论

针对建筑图像检索存在局部遮挡和尺度变化的问题，本文提出了融合全局聚合与局部挖掘的建筑图像检索网络。设计多尺度特征聚合模块获取具有判别性的全局显著特征；提出注意力引导的特征挖掘模块学习更多潜在的局部次显著特征；最后通过特征正交融合策略增强局部特征和全局特征之间的互补性，最大限度地利用提取到的多样化特征。实验结果表明，所提网络 GALM-Net 在 2 个公开的建筑实例检索数据集 ROxf 和 RPar 上的检索精度分别达到了81.54%（M）、62.43%（H）和90.28（M）、

（a）ResNet50 检索结果

(b)GALM-Net 检索结果

图 9　不同网络在相同类别图像上的检索结果

78.35%(H)。与对比网络相比,本文所设计的网络具有较高的检索准确率和鲁棒性。

参考文献

[1] 斯白露, 高文, 卢汉清,等. 基于感兴趣区域的图像检索方法[J]. 高技术通讯,2003,13(5):13-18.

[2] 季长清, 王兵兵, 秦静, 等. 深度特征的实例图像检索算法综述[J]. 计算机科学与探索, 2023,17(7):1565-1575.

[3] DAVID L. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004,60:91-110.

[4] BAY H, ESS A, TUYTELAARS T,et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008,110(3):346-359.

[5] SPYROMITROS-XIOUFIS E,PAPADOPOULOS S,KOMPATSIARIS I Y, et al. A comprehensive study over VLAD and product quantization in large-scale image retrieval[J]. IEEE Transactions on Multimedia, 2014,16(6):1713-1728.

[6] TOLIAS G, JENICEK T, CHUM O. Learning and aggregating deep local descriptors for instance-level recognition[C]//The 16th European Conference on Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020:460-477.

[7] LEE S, SEONG H, LEE S, et al. Correlation verification for image retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022:5374-5384.

[8] GORDO A, ALMAZAN J, REVAUD J, et al. End-to-end learning of deep visual representations for image retrieval[J]. International Journal of Computer Vision, 2017,124(2):237-254.

[9] GU Y, LI C, XIE J. Attention-aware generalized mean pooling for image retrieval[EB/OL]. (2019-02-28)[2023-12-20]. https://arxiv.org/pdf/1811.00202.pdf.

[10] NG T, BALNTAS V, TIAN Y, et al. SOLAR: second-order loss and attention for image retrieval[C]//The 16th European Conference on Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020:253-270.

[11] 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. Venice, Italy: IEEE, 2017:3456-3465.

[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. Long Beach, USA: IEEE, 2019:5109-5118.

[13] CAO B, ARAUJO A, SIM J. Unifying deep local and global features for image search[C]//The 16th European Conference on Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020: 726-743.

[14] 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. Virtual: IEEE, 2021:11772-11781.

[15] SONG C H, HAN H J, AVRITHIS Y. All the attention you need: global-local, spatial-channel attention for image retrieval[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2022:2754-2763.

[16] SONG Y, ZHU R, YANG M, et al. DALG: deep atten-

tive local and global modeling for image retrieval[EB/OL]. (2022-07-01)[2023-12-20]. https://arxiv.org/pdf/2207.00287.pdf.

[17] RADENOVIĆ F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018,41(7):1655-1668.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:770-778.

[19] LIU S, HUANG D, WANG Y. Receptive field block net for accurate and fast object detection[C]//Proceedings of the 15th European Conference on Computer Vision-ECCV 2018. Munich, Germany: Springer, 2018:404-419.

[20] 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.

[21] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2020:11534-11542.

[22] RADENOVIĆ F, ISCEN A, TOLIAS G, et al. Revisiting oxford and paris: large-scale image retrieval bench marking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:5706-5715.

[23] PHILBIN J, CHUM O, ISARD M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007:1-8.

[24] PHILBIN J, CHUM O, ISARD M, et al. Lost in quantization: improving particular object retrieval in large scale image databases[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE, 2008:1-8.

[25] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(8):2011-2023.

[26] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision-ECCV 2018. Munich, Germany: Springer, 2018:3-19.

Fusing global aggregation and local mining for architectural image retrieval

MENG Yuebo, ZHANG Ziqin, LIU Guanghui, XU Shengjun

(College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055)

Abstract

To address the problem of low retrieval accuracy in architectural image retrieval due to scale variations and local occlusions, this paper proposes an architectural image retrieval network that integrates global aggregation and local mining. The method introduces global branch for multi-scale feature aggregation and a local branch for attention-guided feature mining following the ResNet50 backbone network. The network efficiently integrates complementary features from the two branches through an orthogonal fusion module. Specifically, the multi-scale feature aggregation module utilizes mixed dilated convolutions and channel attention to adaptively aggregate globally different-scale targets, enhancing the network's ability to extract multi-scale salient features from architectural images. The attention-guided feature mining module employs information complementary attention to mark and erase the most salient feature, achieving the mining of potential detail information in local regions. The proposed method achieves mean average precision (mAP) metrics of 81.54% (M) and 62.43% (H) on the ROxf dataset, as well as 90.28% (M) and 78.35% (H) on the RPar dataset, which are two major mainstream architectural datasets. Experimental results indicate that the method effectively overcomes the interference of scale variations and local occlusions, significantly improving the accuracy of architectural image retrieval.

Key words: architectural image, image retrieval, feature aggregation, feature mining