基于视觉一致性增强的细粒度图像检索

郎文溪,孙 涵
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要:细粒度图像检索旨在从大类图像中检索出特定子类的图像。得益于卷积神经网络的快速发展,细粒度图像检索的精度和速度均取得突破,但其性能仍受限于不同子类图像间高相似性和同一子类图像间的高差异性。针对上述问题,该文提出了一种基于对比学习和视觉一致性增强的细粒度图像检索框架 CVCS-Net。CVCS-Net 由判别性特征挖掘模块、视觉一致性增强模块和语义哈希编码模块组成,在挖掘类间图像判别性特征的同时,通过增强类内图像的视觉一致性来提升模型对类内图像差异的容忍度。判别性特征挖掘模块学习空间注意力图来定位图像的判别性区域并获得这些区域对应的局部特征表示;视觉一致性增强模块提升模型对类内图像差异的鲁棒性;而语义哈希编码模块基于量化损失和位平衡损失进一步学习紧凑的哈希码用于检索。CVCS-Net 在 CUB200-2011、Stanford Dogs 和 Stanford Cars 的 mAP 分别可达到 0.859 1、0.856 4 和 0.918 3,相较于当前其他检索方法能够取得更好的检索结果。

关键词:细粒度图像检索;弱监督;对比学习;哈希;视觉一致性
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2022)12-0012-09
doi:10.3969/j.issn.1673-629X.2022.12.003

Fine-grained Image Retrieval Based on Strengthened Visual Consistency

LANG Wen-xi, SUN Han
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Fine-grained image retrieval aims at retrieving images of specific sub-categories from general categories of images. Thanks to the rapid development of convolutional neural networks, there has been a breakthrough in the accuracy and speed of fine-grained image retrieval. However, its performance is still limited by the high similarity between images of different sub-categories and the high difference between images of the same sub-category. Therefore, a contrast learning and strengthened visual consistency CVCS-Net is proposed. CVCS-Net consists of three key modules: discriminative feature mining, strengthened visual consistency and semantic hash coding. The discriminative feature mining module learns spatial attention maps to locate discriminative regions of images and obtains local feature representations corresponding to these regions; the strengthened visual consistency module improves the robustness of the model to intra-class image differences; and the semantic hash coding module further learns compact hash codes for retrieval based on quantization loss and bit balance loss. CVCS-Net can get mAPs of 0.859 1, 0.856 4 and 0.918 3 for CUB200-2011, Stanford Dogs and Stanford Cars, respectively, which can get better results compared with other current retrieval methods.

Key words: fine-grained image retrieval; weak supervision; contrast learning; hashing; visual consistency

0 引 言

图像检索[1]一直是计算机视觉领域的热点问题,其基本目标是从海量图像数据库中查询和返回与检索内容相关的图像。随着深度学习的发展,图像检索任务的准确性和速度均取得突破[2-7]。然而,随着用户对搜索引擎检索结果的定制化和精细化,细粒度图像检索近年来逐渐受到学术界和工业界的广泛关注。对于给定的属于同一大类(如,狗)的图像,细粒度图像检索旨在进一步检索属于相同子类的图像(如,沃克猎犬和巴塞特犬)。相较于经典图像检索,细粒度图像检索的主要难点包括:(1)类间差异小。不同子类的图像高度相似,区分性的差异信息仅体现细微的局部区域;(2)类内差异大。相同子类的图像由于姿态、光照、背景和拍摄角度的不同,差异巨大难以区分。因

收稿日期:2021-12-14 **修回日期**:2022-04-19
基金项目:国防科技创新特区项目资助(XX);中央高校基本科研业务费专项资金(NZ2019009)
作者简介:郎文溪(1997-),女,硕士,CCF 会员(B5265G),研究方向为计算机视觉;通讯作者:孙 涵(1978-),男,博士,副教授,CCF 南京秘书长(33361M),研究方向为计算机视觉。

此,将经典图像检索算法应用在细粒度图像数据集[8-11]上效果不佳,区分和检索细粒度图像仍然是目前具有挑战性的研究热点和难点。

细粒度图像较小的类间差异使得不同子类间的差异仅体现在目标局部,而这些细微差异在特征学习时易受目标其他区域的干扰,在最终用于分类和检索的特征图上常被淹没。针对这一问题,一些细粒度研究工作[12-17]致力于挖掘局部区域的判别特征。SCDA[18]使用预训练的 CNN 定位显著前景区域和无关的背景噪声,然后引入 flood-fill 算法过滤噪声获得更具判别性的特征编码。ExchNet[19]和 WSDAN[20]均首先基于注意力机制获取目标的判别性局部区域,然后分别设计通道约束和数据增强策略进一步增强局部区域特征的判别性。虽然这些工作能够显著提升细粒度图像的识别和检索精度,但都忽略了细粒度任务中的另一个关键问题。如图 1 所示,对于姿态、光照、成像角度等因素变化而带来的巨大类内差异,属于相同子类的图像呈现出截然不同的像素分布。在只有图像类别标签时,卷积神经网络在训练过程中很难捕捉到这些不同因素间等变的通用模式。因此,模型对类内图像的视觉变换非常敏感,相同子类图像的识别和检索精度在收敛趋势差异巨大,严重制约了细粒度检索在实际应用中的性能。对于客观存在无法消除的类内图像变化,基于数据驱动的策略,一个简单的解决思路是增强训练数据中任一类别样本的数量。然而,搜集场景丰富的足够训练样本费时费力,且特征学习时模型依然无法高效地动态自适应这些类内变换。针对上述问题,本研究基于对比学习来增强类内图像的视觉一致性,提高模型对视觉变化的容忍度来提升最终的检索精度。具体而言,视觉一致性由两个隐式的正则化约束实现,其基本要求为:(1)同一子类不同图像的局部判别区域在特征空间中应该有相似的表示;(2)判别性特征应该在同一图像的不同视图间保持语义一致。

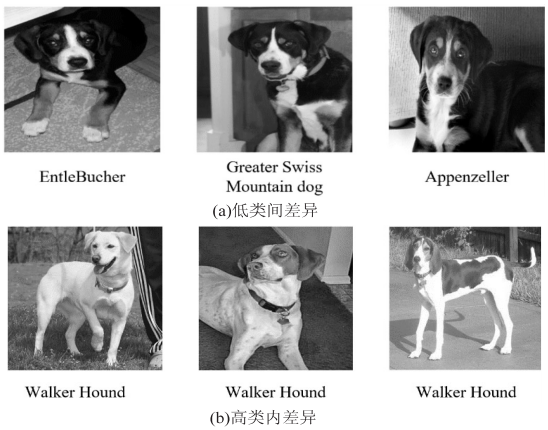


图 1　细粒度图像特点示意

该文提出了一种基于对比学习和视觉一致性增强的细粒度图像检索框架(Contrastive Visual Consistency Strengthen,CVCS-Net),主要包括判别性特征挖掘、视觉一致性增强和语义哈希编码三个关键模块。对于低类间差异,CVCS-Net 设计判别性特征挖掘模块,自适应地学习空间注意力图来定位其与类别最相关的 M 个判别性区域;其次,对于高类间差异,CVCS-Net 受自监督算法[21-22]的启发,设计视觉一致性增强模块提升模型对差异巨大的类内图像的鲁棒性。视觉一致性增强模块由判别区域一致性增强和变换图像一致性增强构成,前者基于语义中心损失来为 M 个判别性区域构造语义一致的通用特征表达,而后者首先应用图像变换策略积极增强类内训练数据的丰富度,其次设计对比度损失惩罚变换后图像的 M 个判别性区域的特征与通用特征表达间的差异。语义哈希编码模块引入量化损失和比特平衡损失,为视觉一致性增强后的判别性特征学习紧凑且高语义的哈希码,最终实现对细粒度图像的准确检索。大量的实验结果表明,提出的 CVCS-Net 能够显著提升细粒度检索精度。主要贡献如下:

(1)提出了一种基于对比学习和视觉一致性增强的细粒度图像检索方法 CVCS-Net,在挖掘判别性特征的同时通过增强网络对细粒度图像类内差异的学习容忍度来提升检索性能。

(2)分别设计语义中心损失和对比度损失来增强细粒度图像的视觉一致性。前者能够引导模型为同一子类的判别性区域构造语义一致的特征中心,而后者能够显著提升模型在检索同一子类差异巨大的不同图像时的鲁棒性。

(3)在三个代表性细粒度数据集上进行了详尽的对比和消融实验,验证了 CVCS-Net 的有效性。

1　相关工作

1.1　细粒度图像检索

早期的细粒度图像检索方法依赖于人工特征的使用[23]。然而人工特征在设计时依赖于大量精确的人工标注信息,无法满足仅使用图像标签作为监督信息的弱监督检索的精度需求。得益于深度学习的迅速发展,近年来越来越多的细粒度检索方法被提出。现有的工作致力于挖掘判别性特征来解决细粒度类间图像差异较小这一问题,主要包括有监督方法和无监督方法两类。SCDA[18]提出了一种无监督的选择性卷积描述符聚合方法,该方法首先定位细粒度图像中的对象,并保留有用的深度描述符以进行细粒度图像检索。对于有监督检索方法,CRL-WSL[24]使用中心排序损失和显著区域轮廓来学习目标的辨别特征,提出一个统一的细粒度检索框架。DCL-NC[25]进一步添加归一

化尺度层和去相关排序损失来改进 CRL-WSL。不同于已有工作,该文在挖掘判别性特征的基础上进一步考虑视觉一致性。挖掘判别性特征有利于为不同子类间高相似性的细粒度目标构建判别性的特征表达,从而实现准确检索;而视觉一致性通过增强网络对不同视觉变换的学习耐受力来解决类内图像差异较大这一问题。

1.2 对比学习

对比学习(Contrastive Learning,CL)[26]在无监督表示学习领域表现出巨大的潜力,其基本动机是使用 InfoNCE loss[27]来估计模型从一组无关负样本中正确分类目标特征表示的能力。对比学习算法设计的关键是高质量正、负样本的构造。khoslet 等[28]设计监督对比损失(SupCon)拉近特征空间中的同类特征之间的距离,同时拉远不同类特征的距离。Li 等[29]提出了 ProtoNCE 损失,利用基于聚类的无监督表示方法促进对比学习精度。最近,Wang 等[30]提出了一种基于像素级的密集对比学习方法,并将其用于多个视觉任务的预训练中。该文将对比学习引入细粒度图像检索任务中,设计一种新的对比度损失来增强细粒度图像的视觉一致性。

1.3 哈希编码

在大规模图像检索任务中,为判别性特征构建哈希编码被证明是高效的解决方案。通过哈希编码,属于相同或不同类别细粒度图像的判别性特征能够被嵌入到相同或不同的二进制码中,在降低存储成本的同时提升了查询速度。现有的哈希方法主要包括 data-independent 哈希和 data-dependent 哈希两类。前者通过随机投影或手工构建二进制哈希码,代表性工作是局部敏感的哈希函数[31](Locally Sensitive Hash,LSH)。相较于直接使用学习到的高维度判别性特征进行检索,LSH 方法虽然提升了检索速度,但依然需要较长的哈希编码来保证检索性能。不同于 data-independent 哈希方法,data-dependent 哈希方法充分挖掘数据内部的隐式内在联系自适应的学习哈希码。根据是否有额外的监督信息,data-dependent 哈希方法又可以分为无监督和有监督的两类。尽管无监督方法[32-34]具有更实用和更广泛的应用前景,但目前绝大部分工作更多的基于有监督的框架,充分利用监督信息来获得更好的检索性能。代表性的有监督哈希算法包括卷积神经网络散列(CNNH)[35]和深度成对监督散列(DPSH)[36]两种。此外,最近的工作中,HashNet[37]提出用 tanh 激活函数不断逼近目标,大大提高了检索性能;DCH[38]提出了一种基于柯西分布的成对交叉熵损失,惩罚汉明距离大于给定阈值的相似图像对来学习哈希码。

2 方 法

该文提出了一种基于对比学习和视觉一致性增强的细粒度图像检索方法,能够实现端到端的模型训练。如图 2 所示,它由三个核心模块组成:判别性特征挖掘模块、视觉一致性增强模块和哈希编码模块。实现不同子类间图像的准确检索是细粒度图像检索任务的基础,为此,提出的 CVCS-Net 首先设计判别性特征挖掘模块,基于空间注意力机制定位目标的关键局部来捕捉不同类别细粒度图像间的差异性特征。其次,CVCS-Net 设计视觉一致性增强模块提升模型对同一子类不同图像间显著差异的容忍度。视觉一致性增强模块由判别区域一致性增强和变换图像一致性增强构成,前者基于语义中心损失来为 M 个判别性区域构造语义一致的通用特征表达,而后者对输入图像随机进行颜色或空间变换,设计对比度损失来惩罚图像变换

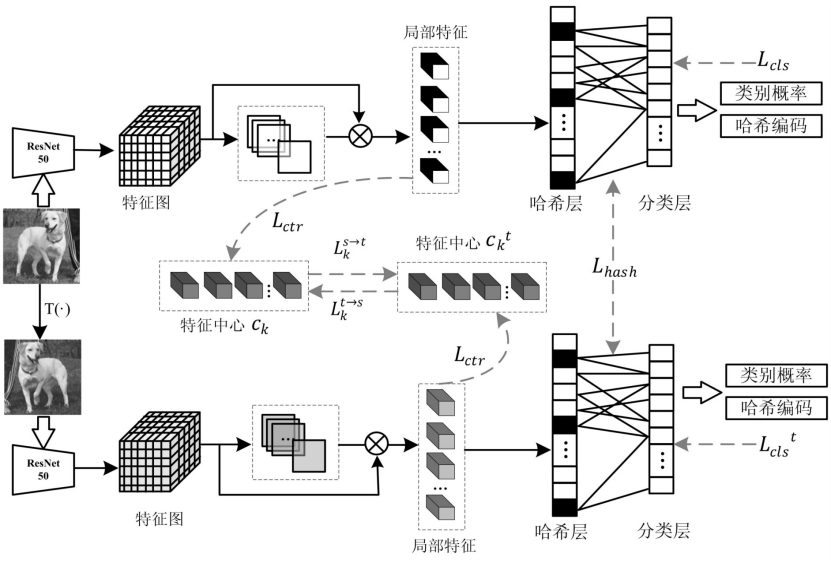


图 2 CVCS-Net 算法流程

前后的特征差异。CVCS-Net 为视觉一致性增强后的判别性特征学习紧凑的语义哈希码实现最终的检索。

2.1　判别特征挖掘

对于细粒度图像分析,较小的类间差异使得捕捉不同类别间的判别性区域至关重要。在弱监督细粒度图像检索任务中,由于模型训练和测试时没有判别性区域对应的标注信息(如:bounding box),判别性特征挖掘模块通过计算目标不同局部的类别得分来学习判别性区域的空间注意力分布图;通过进一步抽取判别性区域对应的局部特征,最终为类内差异小的细粒度图像构造精细的特征表达。

2.1.1　判别区域定位

对于给定的任一训练图像 X,通过特征提取网络首先生成一组特征图 $F \in R^{H\times W\times N}$。其中,$H$、$W$ 和 N 分别为特征图的长、宽和通道数。其次,特征图 F 被输入到一个额外的卷积模块中生成注意力图 $A \in R^{H\times W\times M}$,如公式(1)所示。

$$A = f(F) = \cup_{k=1}^{M} A_k \tag{1}$$

其中,$f(\,)$ 表示卷积操作,由一个卷积核大小为 1 × 1 的卷积层和一个 Relu 激活层实现;$A_k \in R^{H\times W}$ 表示图像的第 k 个注意力图,描述了目标第 k 个局部区域的空间位置信息。

2.1.2　判别特征挖掘

进一步使用注意力图 A 计算这些局部区域对应的特征,计算方法如公式(2)所示:

$$c_k = g(A_k \odot F), k = 1,2,\cdots,M \tag{2}$$

其中,f_k 是第 k 个局部区域对应的特征,$\odot$ 表示将特征图 F 和第 k 个注意力图对应位置元素相乘,$g(\,)$ 表示全局平均池化操作。为了保证局部区域的判别性,这里简单却有效的假设是,如果一个局部区域是判别性的,那么其对应的特征 f_k 在分类时一定能够高度响应其类别。因此,直接使用获得的 M 个局部特征预测目标的类别,计算方法如公式(3)和(4)所示:

$$P = \mathrm{softmax}(f) = \mathrm{softmax}\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}, k = 1,2,\cdots,M \tag{3}$$

$$L_{\mathrm{cls}} = \mathrm{CE}(P,Y) \tag{4}$$

P 表示 softmax 分类器的预测结果。在类别标签的监督下,使用交叉熵损失 CE ( ) 来约束预测结果 P 与其对应的真实类别 Y 趋向于一致。通过上述过程,模型不断寻找与类别最相关的 M 个局部区域,最终捕捉细粒度图像判别性特征。

2.2　视觉一致性增强

2.2.1　判别区域一致性增强

如图 1(b)所示,受光照、姿态、成像角度等诸多因素的影响,相同子类的不同目标差异巨大。目标固有的这些多样性无法消除,对细粒度图像分析带来了极大的挑战。这里,本研究的解决思路是首先通过不同的图像变换模拟多样性,然后基于对比学习的思想增强模型对这些多样性变化的耐受性(称之为视觉一致性)。如图 2 所示,首先对原始图像 X 进行视觉变换,如公式(5)所示:

$$X^t = T(X) \tag{5}$$

其中,X^t 为变换后的图像,$T(\,)$ 表示不同的视觉变换。本研究中视觉变换主要包括颜色抖动(包括亮度、对比度、饱和度和色调)和空间变换(包括水平翻转、缩放等)两大类。变换后的图像将输入到判别性特征挖掘模块,获得其对应的 M 个判别性特征 f^t 和空间注意力图 A^t。其中,f_k^t 和 A_k^t 分别表示变换后图像的第 k 个判别性特征和空间注意力图。基于公式(3)和(4),计算变换后图像的类别概率 P^t 和分类损失 L_{cls}^t。

在增强视觉一致性时,首先设计语义中心损失 L_{ctr} 来保证相同子类不同图像的相似判别性区域在语义空间中有相近的特征表达(如所有沃克猎犬图像的第 k 个空间注意力图都感知其头部)。具体而言,为每一个子类的 M 个判别性区域分别构造特征中,惩罚属于同一判别性区域的局部特征间的变化,如公式(6)所示:

$$L_{\mathrm{ctr}} = \sum_{k=1}^{M} \| f_k - c_k \|^2 \tag{6}$$

其中,c_k 是 f_k 所对应的特征中心。c_k 初始化的值为 0 且在训练被更新,计算方法如公式(7)所示:

$$c_k = (1-\mu)c_k + \mu f_k \tag{7}$$

其中,μ 表示更新权重,该文设置为 0.5。对应变换后的图像,也同样为每一个判别性区域计算特征中心 c_k 和 c_k^t。

2.2.2　变换图像一致性增强

如图 2 所示,为了提高模型对同一子类的不同图像的检索精度,还设计对比度损失进一步增强图像变换前后的视觉一致性。具体而言,CVCS-Net 惩罚 X 和 $T(X)$ 在判别性特征中心上的差异来约束变换前后的图像得到相似的分类和检索结果。对于任一 X 和 $T(X)$,其变换前后的特征中心为 c_k 和 c_k^t,取变换前后的相同特征中心作为正样本对 $(c_k, c_m^t)_{k=m}^{k\in\{1,2,\cdots,M\}}$ 并在编码空间中拉近两者距离;取变换前后不同的特征中心作为负样本对 $(c_k, c_m^t)_{k\neq m}^{k\in\{1,2,\cdots,M\}}$ 并在编码空间中拉远两者距离。首先,变换前的特征中心作为变换后特征中心的监督,第 k 个特征中心上 (i,j) 处的对比损失 $L_k^{s\to t}(i,j)$ 计算方法如公式(8)所示:

$$L_k^{s\to t}(i,j) = -\log \frac{\exp(c_k(i,j)\cdot c_k^t(i,j)/\varphi)}{\sum_{m=1,m\neq k}^{M} \exp(c_k(i,j)\cdot c_m^t(i,j)/\varphi)}$$

$$(i,j) \in c_k \tag{8}$$

其中，φ 是平衡系数，本研究中设置为 0.1。类似的，也将变换后的特征中心作为变换前特征中心的减速，此时第 k 个特征中心上 (i,j) 处的对比损失 $L_k^{t\to s}(i,j)$ 计算方法如公式(9)所示：

$$L_k^{t\to s}(i,j) = -\log \frac{\exp(.\, c_k^t(i,j).\, c_k(i,j)/\varphi)}{\sum_{m=1,m\neq k}^{M} \exp(c_k^t(i,j).\, c_m(i,j)/\varphi)} \quad (i,j) \in c_k \tag{9}$$

对比度损失 L_{con} 计算 M 个特征中心上的所有像素点，如公式(10)所示：

$$L_{\text{con}} = \frac{1}{H \times W} \sum_{k=1}^{M} \sum_{i=0,j=0}^{i=H,j=W} (L_k^{s\to t}(i,j) + L_k^{t\to s}(i,j)) \tag{10}$$

2.3 语义哈希编码

在大多数深度哈希方法中，哈希层被设计来将目标特征编码为二进制的哈希码，其中“1”表示类别拥有某种特征，而“-1”表示它缺乏这种特征。CVCS-Net 在分类层前设计并添加了一个语义哈希编码模块，将视觉增强后的特征 f^t 映射为 B 位的哈希码，其基本计算方法如公式(11)所示：

$$H_i = \tanh((W^H)^{\mathrm{T}} f_i^t + \delta^H),\ i = 1,2,\cdots,M \tag{11}$$

其中，$H_i \in R^B$ 是 f^t 经过哈希层的输出；$\delta^H \in B$ 和 $W^H \in R^{M\times B}$ 分别表示哈希层的偏差和权重；tanh( ) 为激活函数，计算方法如公式(12)所示：

$$\tanh(x) = \frac{e^{f} - e^{-f}}{e^{f} + e^{-f}} \tag{12}$$

tanh( ) 的取值范围为[-1,1]。由于 sign 函数在非零点处的梯度可能为零，出现梯度消失的问题，因此只有在测试模型时，将实值输出的 H_i 映射为二维哈希码。若 $H_i \geqslant 0$，则 $B_i = 1$；否则，$B_i = -1$。

$$B_i = \text{sign}(H_i) \tag{13}$$

在优化 sign 函数时，为了避免梯度消失，CVCS-Net 在训练阶段松弛了二进制哈希码。而对于测试阶段，将实值特征表示转换为二进制化的哈希码可能会导致量化误差。在这种情况下，添加量化损失 L_q 以鼓励实值尽可能地接近所期望的哈希码，如公式(14)所示。

$$L_q = \sum_{i=1}^{N} (|H_i| - e)^2 \tag{14}$$

其中，e 是所有元素值均为 1 的 B 维向量，B 为哈希码的长度。此外，还增加额外的位平衡损失 L_b，以使哈希码的每一个比特位为 1 或-1 的可能性相等，如公式(15)所示：

$$L_b = \sum_{i=1}^{N} \text{mean}(H_i) \tag{15}$$

其中，mean( ) 的计算方法如下：

$$\text{mean}(H) = \frac{1}{B} \sum_{n=1}^{K} \text{mean}(H_n) \tag{16}$$

其语义编码模块损失为：

$$L_{\text{hash}} = L_q + L_b \tag{17}$$

整体的损失函数如公式(18)所示：

$$L = L_{\text{cls}} + L_{\text{cls}}^t + L_{\text{ctr}} + L_{\text{con}} + L_q + L_b \tag{18}$$

3 实验结果与分析

3.1 数据集及评价指标

为了验证 CVCS-Net 的有效性，在 CUB Birds、Stanford Cars 和 Stanford Dogs 三个细粒度数据集上进行了实验。CUB Birds 包含 11 788 张图像，共 200 个类别。数据集包括 5 794 张图像的训练集和 5 994 张图像的测试集。Stanford Cars 数据集由 8 144 张训练图像和 8 041 张测试图像组成，属于 196 个类别。Stanford Dogs 数据集由 120 个类别的 20 580 张图像组成，包括 12 000 张图像的训练集和 8 580 张图像的测试集。

使用平均检索精度(mAP)定量评估检索性能，计算方法如公式(19)所示：

$$\text{mAP} = \frac{1}{n_q} \sum_{i=1}^{n_q} \text{AP}, \text{AP} = \frac{1}{n} \sum_{j=1}^{K} \frac{n_j}{j} \text{pos}(j) \tag{19}$$

其中，n_q 是需要检索的样本数；n 是每次检索后返回的样本数，n_j 是返回的样本 n 中的正样本数量。如果返回的第 j 张是正样本，则 pos(j) 值为 1，否则值为 0。

3.2 参数设置

实验中，基于 pytorch 实现代码并使用 CPU Intel i5，GPU RTX 2080ti 11 GB，32 G 内存的硬件平台训练模型。缩放训练图像的尺寸为 448 × 448 像素，并使用 ResNet-50 作为主干来提取特征并选择 Conv5 层的输出作为特征图。对于注意力图的生成，M 的默认值为 32。

尽管所有数据集都标有边界框或零件位置，但提出的 CVCS-Net 仅使用类别标签作为监督信息。在训练阶段，模型使用随机梯度下降(SGD)训练 160 个 epoch，batch size 为 12，初始学习率设置为 0.001 且每 2 个 epoch 后进行指数衰减，衰减系数为 0.9。在推理阶段，遵循文献[39]中的设定，使用测试图像作为查询集，训练图像作为所有实验的检索数据库。

3.3 细粒度检索性能对比分析

3.3.1 与传统图像检索方法性能比较

为了验证该方法的有效性，首先比较了 CVCS-Net 与其他传统图像检索方法在细粒度数据集上的检索精度。为了保证比较的公平，对所有方法均使用 Resnet-50 作为特征提取网络，并使用不同长度的哈希码分别训练模型。

如表 1 所示,对于 CUB 数据集,CVCS-Net 在 12 位、32 位和 48 位哈希码时的 mAP 分别为 79.23%、84.69% 和 85.91%,较性能第二的 FPH 提升均超 24%。对于 Dogs 和 Cars 数据集,CVCS-Net 提升的趋势依然明显。上述结果一方面表明了细粒度图像检索任务与普通图像检索任务的巨大差异,另一方面也证明了判别性特征挖掘对细粒度图像分析任务的重要性。此外,相较于使用全局图像特征构建哈希码的经典检索方法,CVCS-Net 对哈希码长度的变化表现出更好的稳定性。例如,当哈希码从 48 位降低到 12 位时,在 CUB 200-2011 数据集上 CVCS-Net 的 mAP 仅下降约 6%,远小于 FPH 的 11%。这一结果充分表明 CVCS-Net 设计判别性特征挖掘模块的有效性,也从侧面验证了语义哈希编码模块在量化损失和位平衡损失的约束下能够提升哈希码的紧凑性和语义性。

表 1　CVCS-Net 与传统图像检索方法在不同数据集上的 mAP 比较

方法	CUB 200-2011			Stanford Dogs			Stanford Cars		
	12bit	32bit	48bit	12bit	32bit	48bit	12bit	32bit	48bit
DHN	0.371 1	0.417 2	0.460 2	0.455 9	0.529	0.573 6	0.460 8	0.505	0.557 4
DQN	0.378 9	0.435 5	0.481 1	0.467 6	0.523 4	0.579 5	0.489 7	0.544 4	0.582 1
HashNet	0.402 7	0.471 2	0.510 3	0.498 8	0.557 4	0.598 1	0.507 3	0.550 8	0.583 2
DCH	0.460 2	0.523 3	0.574	0.608 1	0.656 7	0.677 9	0.548 8	0.600 9	0.617 5
FPH	0.512 8	0.583 2	0.617 9	0.631 2	0.690 9	0.709	/	/	/
Ours	0.792 3	0.846 9	0.859 1	0.772 0	0.834 9	0.856 4	0.904 4	0.910 8	0.918 3

3.3.2　与细粒度检索方法性能比较

进一步的,还比较了 CVCS-Net 与已有细粒度图像检索方法在 CUB 数据集上的检索精度。如表 2 所示,CVCS-Net 在不同设定下均能够取得最佳的检索性能。

表 2　CVCS-Net 与细粒度图像检索方法在 CUB 数据集上的 mAP 比较

方法	特征提取网络	特征维度	mAP
SCDA	VGG16	4 096	0.595 7
CRL WSL	VGG16	1 024	0.659
DCL-NS	Resnet-50	1 024	0.679
ExchNet	Resnet-50	12	0.251 4
		32	0.677 4
		48	0.710 5
FCAENet	Resnet-50	12	0.347 6
		32	0.738 5
		48	0.801 4
文献[41]	Resnet-50	12	0.790 1
		32	0.842 6
		48	0.854 9
CVCS-Net	Resnet-50	12	0.792 3
		32	0.846 9
		48	0.859 1

相较于没有哈希模块的细粒度检索方法 SCDA、CRL-WSL 和 DCL-NS,即使其特征表示的维度远大于 CVCS-Net,其精度仍然远低于提出的 CVCS-Net。可能的原因是细粒度任务对特征质量的敏感。虽然高维图像特征包含了更多的物体信息,但在计算不同图像的相似度时,判别性特征可能无法起主导作用,使得最终检索结果更容易被相似的类间图像干扰。相较于基于哈希的细粒度检索方法 ExchNet 和 FCAENet[40],在哈希码长度为 48 位时,CVCS-Net 的 mAP 为 85.91%,分别提升约 14% 和 5%。此外,ExchNet 和 FCAENet 的性能还随着哈希码长度的变化波动剧烈。当哈希码从 12 位增加到 32 位时,这两种方法的 mAP 提高约 40%,而提出的 CVCS-Net 仅提升 6%。当哈希码从 32 位增加到 48 位时,CVCS-Net 的 mAP 仅增加约 1%。上述的实验结果表明,视觉一致性增强能够显著提升模型的鲁棒性。

3.4　消融实验

3.4.1　不同模块有效性验证

如表 3 所示,该实验进一步验证了 CVCS-Net 每个部分的有效性。以 48 位哈希码为例,CVCS-Net 简单的基线是仅使用判别性特征和语义哈希编码模块,其 mAP 为 80.84%。在分别加入语义中心损失和对比度损失进行视觉一致性增强后,mAP 能够提升 2.05% 和 3.42%;而 L_{ctr} 和 L_{con} 共同作用下 mAP 为 84.62%。上述的结果充分验证了语义中心损失和对比度损失的有效性。首先,通过构建语义中心,CVCS-Net 能充分挖掘类内不同目标的相似判别性区域;其次,对比度损失在训练时能够极大地提高模型对类内图像变化的耐受性。类似的,当基线加入量化损失 L_q 和位平衡损失 L_b 后,其结果能够从 80.84% 分别提升到 82.46% 和 81.58%。实验结果表明,量化损失和位平衡损失能够显著提升语义哈希编码的质量和检索性能。

表 3 CUB 数据集上不同设置时 mAP 的比较

判别性特征挖掘模块	视觉一致性增强模块		语义哈希编码模块		哈希位数	
	L_{ctr}	L_{con}	L_q	L_b	12	48
√					0.719 3	0.808 4
√	√				0.736 1	0.828 9
√		√			0.753 7	0.842 6
√	√	√			0.760 2	0.846 2
√			√		0.731 5	0.824 6
√				√	0.725 3	0.815 8
√			√	√	0.747 6	0.831 7
√	√	√	√	√	0.792 3	0.859 1

3.4.2 判别性特征数量分析

对细粒度检索任务而言,判别性特征的数量将显著影响模型的性能。如表 4 所示,本小节在不同的细粒度数据集上进行了详尽的消融实验来分析判别性特征数据对 CVCS-Net 的性能影响。以最具挑战性的 CUB 200-2011 数据集为例,在简单挖掘 4 个判别性特征时,CVCS-Net 的 mAP 为 84.11%。这一结果随着特征数量的增加而不断提升,在特征数量为 32 和 64 时 mAP 分别为 85.91% 和 86.52%。类似的上升趋势也显示在其他两个数据集上。在实验中,考虑计算开销和检索性能之间的平衡,最终选择将 CVCS-Net 挖掘的判别特征数设置为 32。

表 4 不同数据集上判别性特征数量与 CVCS-Net 结果比较

判别性特征数量	数据集		
	CUB 200-2011	Stanford Dogs	Stanford Cars
4	0.841 1	0.839 6	0.900 7
8	0.850 7	0.847 8	0.909 2
16	0.855 9	0.854 5	0.917 3
32	0.859 1	0.856 4	0.918 3
64	0.863 2	0.860 1	0.919 5

3.5 可视化效果分析

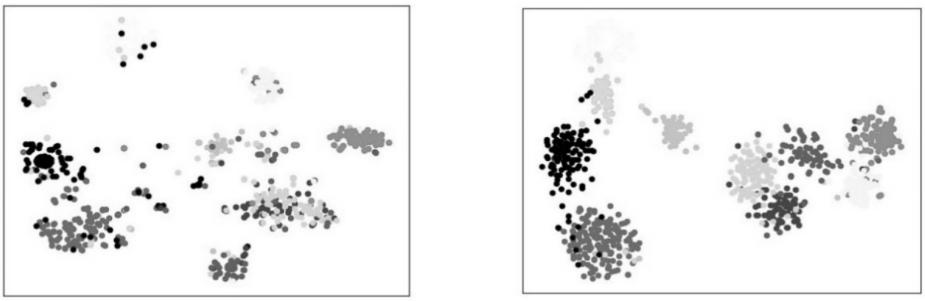
3.5.1 检索效果可视化

图 3 显示了 CVCS-Net 和 HashNet 在三个细粒度数据集上检索的前 10 个样本。可视化的结果表明,CVCS-Net 能够取得比目前最先进方法更满足用户期望的检索结果。

图 3 不同数据集上 CSCV-Net 与 HashNet 检索结果对比

3.5.2　哈希码边界可视化

为了分析哈希码的质量,在 Stanford Cars 数据集上使用 t-SNE(t-distributed stochastic neighbor embedding)将生成的哈希码可视化。如图 4 所示,随机采样 10 个类别的结果表明,CSCV-Net 得益于为细粒度对象特别设计的判别性特征提取模块和语义哈希编码模块,能够将哈希码在同类中生成的更清晰和紧凑,在不同子类间生成的边界更明显。



(a)HashNet　　(b)CVCS-net

图 4　Stanford Cars 上 CSCV-Net 与 HashNet 哈希码可视化对比

4　结束语

针对细粒度图像类间差异小类内差异大的特点,在挖掘判别性特征的基础上基于对比学习进一步增强不同图像间的视觉一致性,提出了一种新颖的深度哈希细粒度图像检索方法 CSCV-Net。CSCV-Net 由判别性特征挖掘模块、视觉一致性增强模块和语义哈希编码模块构成。判别性特征挖掘模块学习空间注意力图,能够有效挖掘图像的判别性局部区域;而视觉一致性增强模块分别引入语义中心损失和对比度损失,增强模型训练时对不同类内图像相似判别区域语义一致性和对同一图像不同视觉变换的鲁棒性;CSCV-Net 还基于量化损失和位平衡损失设计语义哈希编码模块,进一步提升检索速度。通过大量实验和消融研究,在 3 个常用细粒度图像数据集 CUB-200-2011、Stanford Cars 和 Stanford Dogs 验证了 CVCS-Net 的有效性。相较于当前其他检索方法,CVCS-Net 能够取得更好的检索结果。

参考文献:

[1] ZHOU W,LU Y,LI H,et al. Spatial coding for large scale partial-duplicate web image search[C]//Proceedings of the 18th ACM international conference on multimedia. Firenze: ACM,2010:511-520.

[2] 刘　颖,程　美,王富平,等. 深度哈希图像检索方法综述[J]. 中国图象图形学报,2020,25(7):1296-1317.

[3] 万　方,强浩鹏,雷光波. 自监督深度离散哈希图像检索[J]. 中国图象图形学报,2021,26(11):2659-2269.

[4] 张　顺,龚怡宏,王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报,2019,42(3):453-482.

[5] 柯圣财,赵永威,李弼程,等. 基于卷积神经网络和监督核哈希的图像检索方法[J]. 电子学报,2017,45(1):157-163.

[6] LIN K,YANG F,WANG Q,et al. Adversarial learning for fine-grained image search[C]//2019 IEEE international conference on multimedia and expo (ICME). Shanghai: IEEE,2019:490-495.

[7] 任夏荔,陈光喜,曹建收,等. 基于深度学习特征的图像检索方法[J]. 计算机工程与设计,2018,39(2):503-510.

[8] WAH C, BRANSON S, WELINDER P, et al. The CaltechUCSD Birds200-2011 dataset[J]. Advances in Water Resources,2011,24(6):1227-1234.

[9] KHOSLA A,JAYADEVAPRAKASH N,YAO B,et al. Novel dataset for fine-grained image categorization[C]//First workshop on fine-grained visual categorization. [s. l.]:[s. n.],2011:2434-2445.

[10] KRAUSE J,STARK M,DENG J,et al. 3d object representations for fine-grained categorization[C]//Proceedings of the IEEE international conference on computer vision workshops. Sydney:IEEE,2013:554-561.

[11] MAJI S,RAHTU E,KANNALA J,et al. Fine-grained visual classification of aircraft[J]. arXiv:1306.5151,2013.

[12] 范业嘉,孙　涵. 基于轻量级深度哈希网络的细粒度图像检索[J]. 计算机技术与发展,2021,31(10):128-133.

[13] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. Santiago:IEEE,2015:1449-1457.

[14] LI P,XIE J,WANG Q,et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE,2018:947-955.

[15] YANG Z,LUO T,WANG D,et al. Learning to navigate for fine-grained classification[C]//Proceedings of the European conference on computer vision (ECCV). Salt Lake City: Springer,2018:420-435.

[16] ZHENG H,FU J,MEI T,et al. Learning multi-attention convolutional neural network for fine-grained image recognition

[C]//Proceedings of the IEEE international conference on computer vision. Venice:IEEE,2017:5209-5217.
[17] FU J,ZHENG H,MEI T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE, 2017:4438-4446.
[18] CUI Q,JIANG Q Y,WEI X S,et al. ExchNet:a unified hashing network for large-scale fine-grained image retrieval [C]//European conference on computer vision. Glasgow: Springer,2020:189-205.
[19] 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.
[20] HU T,QI H,HUANG Q,et al. See better before looking closer:weakly supervised data augmentation network for fine-grained visual classification[J]. arXiv:190109891,2019.
[21] HE K,FAN H,WU Y,et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle:IEEE,2020:9729-9738.
[22] WU Z,XIONG Y,YU S X,et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City:IEEE,2018:3733-3742.
[23] XIE L,WANG J,ZHANG B,et al. Fine-grained image search[J]. IEEE Transactions on Multimedia,2015,17(5): 636-647.
[24] ZHENG X,JI R,SUN X,et al. Centralized ranking loss with weakly supervised localization for fine-grained object retrieval[C]//Proceedings of the twenty seventh international joint conference on artificial intelligence. Sweden: IJCAI, 2018:1226-1233.
[25] ZHENG X,JI R,SUN X,et al. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer[C]//Proceedings of the AAAI conference on artificial intelligence. Honolulu: AAAI Press, 2019: 9291-9298.
[26] JAISWAL A,RAMESH B A,ZAKI Z M,et al. A survey on contrastive self-supervised learning[J]. Machine Learning, 2020,12:4182-4192.
[27] OORD A V D,LI Y,VINYALS O. Representation learning with contrastive predictive coding [J]. arXiv: 180703748, 2018.
[28] KHOSLA P, TETERWAK P, WANG C,et al. Supervised contrastive learning[J]. arXiv:200411362,2020.
[29] LI J,ZHOU P,XIONG C,et al. Prototypical contrastive learning of unsupervised representations [J]. arXiv: 200504966,2020.
[30] WANG X, ZHANG R, SHEN C,et al. Dense contrastive learning for self-supervised visual pre-training[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville:IEEE,2021:3024-3033.
[31] DATAR M,IMMORLICA N,INDYK P,et al. Locality-sensitive hashing scheme based on p-stable distributions[C]// Proceedings of the twentieth annual symposium on computational geometry. Brooklyn:ACM,2004:253-262.
[32] GONG Y,LAZEBNIK S,GORDO A,et al. Iterative quantization:a procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2013,35(12):2916-2929.
[33] LIU H,JI R,WU Y,et al. Towards optimal binary code learning via ordinal embedding [C]//Proceedings of the AAAI conference on artificial intelligence. Phoenix: AAAI Press,2016.
[34] 郑筱智,李景华. 医学口腔图像检索的快速无监督多模态哈希方法[J]. 信息技术与信息化,2021(6):123-125.
[35] XIA R,PAN Y,LAI H,et al. Supervised hashing for image retrieval via image representation learning[C]//Twenty-eighth AAAI conference on artificial intelligence. Québec: AAAI,2014.
[36] LAI H,PAN Y,LIU Y,et al. Simultaneous feature learning and hash coding with deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston:IEEE,2015:3270-3278.
[37] CAO Z,LONG M,WANG J,et al. Hashnet:deep learning to hash by continuation[C]//Proceedings of the IEEE international conference on computer vision. Venice: IEEE, 2017: 5608-5617.
[38] ZHU H,LONG M,WANG J,et al. Deep hashing network for efficient similarity retrieval[C]//Proceedings of the AAAI conference on artificial intelligence. Phoenix: AAAI Press, 2016.
[39] YANG Y,GENG L,LAI H,et al. Feature pyramid hashing [C]//Proceedings of the 2019 on international conference on multimedia retrieval. Ottawa:ACM,2019:114-122.
[40] ZHAO Q,WANG X,LYU S,et al. A feature consistency driven attention erasing network for fine-grained image retrieval[J]. arXiv:211004479,2021.
[41] SUN H,FAN Y,SHEN J,et al. A novel semantics-preserving hashing for fine-grained image retrieval[J]. IEEE Access,2020,8:26199-26209.