指导教师：杨涛　　提交时间：2016.3.19

CVPR2015 Paper Translation

No: 01
姓名：杨涛
学号：2013302585
班号：10011305

用一个例子进行属性和类别泛型实例搜索

Ran Tao1, Arnold W.M. Smeulders1, Shih-Fu Chang2

ISLA, Informatics Institute, University of Amsterdam, The Netherlands

Department of Electrical Engineering, Columbia University, USA

介绍摘要

本文旨在从一个泛型实例搜索例子，该实例可以是任意的三维对象例如鞋子，而不只是近似平面的和片面的例子如建筑和标志。首先，我们评估在这个问题上最先进的搜索方法。我们观察到，是什么在起作用使建筑物失去与鞋子的共性。其次，我们建议使用自动学习类别特定的属性，以解决目前在泛型实例中搜索大的外观变化。在这个问题来自同一类别的实例之间的搜索作为查询，特定类别的属性胜过大多数现有的方法。在一个包含 6624 张全视角鞋子图片资料组中，我们用特定类别属性（这种方法）将搜索性能从 36.73 改善到了 56.56。第三，我们扩展我们的方法来搜索对象，而不限制于具体已知的范畴。我们展示类级别信息的组合和对特定类别的属性是优越于有低级别功能（如费舍尔矢量合成）的类级信息。

1. 简介

在实例搜索，目的是检索给定的该对象[3，16，25，31]的几个查询示例的特定对象的所有图像。我们只考虑 1 查询图像的具有挑战性的情况下，并在承认差异大

成像角和查询图像和目标图像之间的其他成像条件。一个非常艰苦的情况下是正面视图中指定的查询，而在搜索设置相关的图像显示从从未背视图之前已经见过。人类通过使用两种类型的一般知识解决搜索任务。首先，当查询实例是某一类，比方说女性，答案须限制为来自同一类。并且，在示出一个属性的正面视图查询，说棕色头发，将限制答案显示相同的或没有这样的属性，即使当视点是从背面。在实例搜索，目的是检索所有图像特定对象的特定该对象[3，16，25，31]的几个查询示例。我们只考虑 1 查询图像的具有挑战性的情况下，并承认在成像角之间的巨大差异和其他在查询图像和目标图像之间的成像条件。一种非常困难的情况是一个查询指定在额叶视图，而相关图像搜索集中展示了从未看到过的背视图。人类通过使用两种类型的一般知识解决搜索任务。首先，当查询实例已确定的类别，比如女性，结果应该被限制为来自同一类。而且，在查询显示一个属性的正面视图，假设棕色的头发，将限制答案显示相同的或者没有这样的属性，即使当视点是来自背面。在本文

中，我们使用并评估这两种类型的一般知识来处理各种各样的实例搜索的情况。

在实例搜索中，已经通过限制搜索建筑物[3，4，29]达到了很好的结果。搜索建筑可以在位置识别和三维重建使用。另一套良好的效果已经达到了寻找标志[20，34，40]来估计商标的外观。以及[44]图书和杂志封面的搜索。所有这些实例搜索的情况下对于近平面的以及片面对象显示了不错的效果。在这项工作中，我们的目标在于扩大到更多级别的实例查询。我们的目标是执行从 1 开始的泛型实例搜索。通用暗示我们考虑任意对象，而不是仅仅平面物体。并且，通用意味着我们的目标是用于特定类型的情况下，例如 RANSAC 的未优化的一种方法匹配的对象的平面性。在我们的案例中，实例不仅可以是建筑物和标识，也可以是鞋，衣服等对象。

在实例中搜索所面临的挑战就是代表/表示查询图像不变量到（未知）外观变化同时维持该查询的足够丰富表示从其他类似的实例允许区分。为了解决这个问题，大多数现有的实例搜索中的方法匹配潜在目标中的斑点的出现对于查询[16，19，29，40，41]。在两个图像之间匹配的质量接近于在所有局部描述符对相似的总和。这之间的差异在于位于局部描述符进行编码的方式的引用的处理方法和在相似性的计算之间。这种模式的建筑，标志和场景的远景上良好的性能得到了实现。然而，在搜索时为与更广泛的范围视点变性，多个侧面，并且可能具有自遮挡和任意事非刚性变形，这些方法都可能作为局部描述符的匹配变得在这些情况下失败。对于只从一个实例实例搜索，总是需要在目标对象上的更多信息，其耦合到对外观变化有力的的表示而成功地解决通用实例搜索。

在本文中，我们建议使用属性的代表性[10，23]，以处理大范围的视觉体现。以这种方式，我们的目标是针对比低级别图像表示外观的变化更坚固，像词袋直方图[39]和 Fisher 矢量[28]。我们依靠属性表示，它已在当训练例子不足以覆盖特征空间[10，45]中的所有变化时显示出了在分类上的优点，可靠地存在于一实施例具有挑战性的情况。我们建议自动学习特定类别的列表和非语义的属性，这在同一类别的实例中有所区别。一个实例可以被表示为一特定组合属性和实例搜索归结为找到属性最相似的组合。

为了处理该查询可能的混淆与其他类别的情况下，我们主张先搜索在概念水平，然后放大到搜索的类别内。以这种方式，我们能够减少所有的像素配置的搜索空间巨大，而仍在有理由相信，我们不会失去目标。

有利的是，当仅存在 1 查询图像，以用稍微用户提供的信息。此外对象区域中的交互规范查询图像，在本文中，我们需要规范该类别查询实例所属。

2. 相关工作

在实例中搜索大多数方法依赖于聚会[29 16 41 40 19]局部图像描述符的比赛中，其中区别驻留在局部描述符进行编码的方式与两个描述符的匹配得分进行评价。一袋字（弓）[39，29]编码本地由最近的视觉词的索引描述符。海明嵌入[16]通过增加一个蝴蝶结后提高额外的二进制代码，以更好地描述在空间中的局部描述符的位置。匹配得分的一对的描述符是 1，如果它们在相同的字和二进制签名之间的汉明距离大于较小一定的阈值。VLAD [17]和 Fisher 矢量[27]由表示与局部描述符改进对弓一个额外的残差向量，通过减去平均得到的的视觉词或分别高斯分量。在 VLAD 和 Fisher 矢量，两个描述符的分数是残差的点积，当他们在相同的字，否则为 0。[41，40]提高 VLAD 和通过用阈值替代点积费希尔矢量多项式的相似性和指数分别相似度给予更多的信贷，以接近描述符对。 [19]仅考虑方向编码局部描述符到视觉词中心，而不是幅度，跑赢上搜索例如费舍尔载体。用这些方法，良好的性能在建筑物上已经实现，标志，和从远处场景。这些实例可以被看作近似平面的和片面的。对于建筑，标志，并从远处的场景中的观察角度的变化被限制为 90 度的象限至多出的 360 圈。为了在有限的观点变化，火柴局部描述符的可之间可靠地建立查询和一个相关的例子。在这项工作中，我们考虑泛型实例搜索，不仅是平面的情况下搜索，在那里实例可以是具有更宽范围的任意物观点变性和多面。我们评估现有对泛型实例实例的这一问题大致平坦例如搜索方法。属性[10，11，23]最近备受关注。它们被用来表示共同的视觉特性的不同的对象。属性表示已使用的图像分类[10，45，2]。属性已所示，当训练实例是是有利充分覆盖在低级特征空间[10，45]的外观的变化。受此启发，我们提出使用属性表示，解决泛型实例搜索，那里只有 1 例如提供有仍然存在着广泛的外观变化。属性已被用于图像检索[38，21，46，45，32]。在[38，21，46]，该查询是由文字属性定义的，而不是图像和目标是返回图像参展查询属性。查询属性需语义含义。在这项工作中，我们要解决的实例进行搜索给定的一个查询图像，这是一个不同的任务，因为正确答案必须表现出相同的情况下，和我们使用非语义属性。 [45，32]也考虑非语义属性，但是对于类别检索，而不是实例搜索。改善实例搜索使用的类别级别的信息已探索在[48，8，13]。 [13]用途类别标签学习投影映射原特征到维下部空间，使得低维度的特征结合某些类别级信息。在这项工作中，代替学习特征映射，我们增加额外特征的原始表示捕获类别级信息。[8]扩展费希尔矢量表示的概念从大的概念，2659 分类器输出向量大规模本体概念多媒体（LSCOM）[24]。在[48]，1000 维概念表示[1]是用来缩小的图像之间的语义一致性的基础

上，倒排索引。既[48]和[8]结合用低级表示类别级信息。在这项工作，我们认为类级别的组合与特定类别的属性信息，而不是低级别表示。

2. 1 贡献

我们的工作提出了三个贡献。我们建议追求从 1 例如泛型实例的搜索中，该实例可以从一个广泛的记录任意 3D 对象范围成像角度。我们证明，这个问题比大约持平的和片面的情况下更难搜索建筑物[29]，徽标[20]和远程场景[16]。

我们评估这个问题上最先进的方法。我们观察到最适合的建筑失去了它的通用性为鞋子和反向是什么在起作用建筑物恶化可以鞋工作。其次，我们建议使用自动学习类别的特定属性来处理范围广泛的泛型实例搜索的外观变化。这里我们假设我们知道查询实例的类别，它提供关键的知识时，只有一个查询图片。查询类别的信息可以给出通过交互式用户界面或自动图像分类（例如，鞋，服饰等）。上的问题来自同一类别的实例中搜索查询特定类别的属性超过已有大幅度的实例的搜索方法，当大的变化的外观存在。作为我们的第三个贡献，我们扩展我们的方法不限制已知类别搜索对象。我们提出，以增加与特定类别属性这是由深进行类别级信息学习从大规模图像分类和类别级分类评分了解到特点。我们显示与特定类别属性优于类别信息低层次的功能，如费舍尔矢量合成类相结合的层次信息。

3. 泛型实例搜索的难度

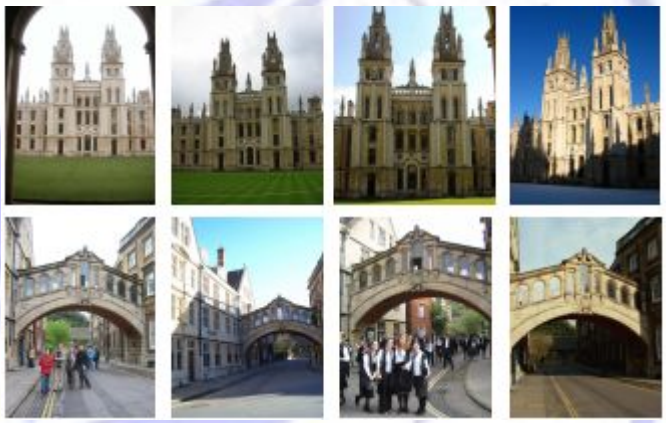
我们在这项工作中提出的第一个问题是最先进的方法如何运用在泛型实例搜索，并且如何执行一个例子在查询实例可以是任意事情。我们可以搜索其他对象，如鞋的建筑物同样的方法？为此，我们对两个数据集进行评估的几个现有的实例搜索算法中，牛津建筑物数据集[29]，收集鞋数据集在这项工作中。 我们评估四种方法。 M1（讲解）：最近工作[40]在两个层面介绍了局部性，以提高从一个例子比如搜索。该方法通过在每个数据库中的图像的评估多个候选位置认为在有关画面局部性。它也通过有效地使用大认为在特征空间局部性对于 VLAD 和 Fisher 矢量和指数相似函数的视觉词汇给予高得不成比例分数上的密切局部描述符对。在当地为实例搜索时的画面显示出有效仅覆盖图像的一部分。而当地的特征空间中示出的所有在本文考虑的数据集是有用的。 M2（Triemb）：[19]提出了三角嵌入和民主的聚集。三角嵌入编码相对于局部描述仅使用方向，而不是震级视觉词中心。如在本文中所示，三角嵌入优于费希尔矢量[35]。民主凝聚分配权重，从提取的每个局部描述图像以确保所有描述同样的贡献图像的自相似性。这种聚合方案是比所示的总和聚集越好。 M3（费舍尔）：我们也可以考虑费舍尔载体，因为它已被

广泛应用其中，良好的业绩有报道实例，搜索和对象分类[18，35]。 M4（深）：有最近显示在顶部层中的激活深卷积神经网络（CNN）[22]作为服务好等特点几个计算机视觉任务[33，6，12]。我们评估对泛型实例的深度学习功能搜索。

数据集。牛津建筑物数据集[29]，通常称为作为牛津大学 5K，包含下载 5062 的图像 Flickr 的。牛津的地标 55 查询所定义，每一个由查询的例子。牛津 5K 是例如搜索，这已被使用的最流行的可用数据集之一许多作品以评估他们的方法。图 1a 示出了从数据集两座建筑物。作为第二个数据集，我们将收集从一组鞋图片亚马逊。它包括 1000 不同的鞋，并在总 6624 的图像。每只鞋从多个观看角度包括来自前，后，顶，底视图记录、一边和其他一些人。鞋的一个形象被认为是作为查询和目标是检索所有其它图像的相同的鞋。虽然这些图像是用干净背景经常看到的购物网站，这是一个有挑战性数据集主要是由于相当大的观点出发，变型和自遮挡的存在。我们将此数据集作为干净鞋子。图 1b 示出了 3 鞋从“干净鞋子”。有一个鞋数据集可用，由[7]提出。然而，该数据集是不适合例如搜索，因为它不包含多个图像一鞋。 [37]还考虑鞋的图像，但该图像完全一致，而在“干净鞋子”图像提供更广泛的观点的变化。

实现细节。对于 M1，M2 和 M3，我们使用黑森州仿射探测器[26]提取兴趣点。在 SIFT 描述符变成 RootSIFT [4]。该充分 128D 描述用于 M1 和 M2，下述[40，19]，而对于费舍尔矢量，局部描述符是减少使用 PCA 到 64D，作为 PCA 减少已显示了费舍尔的重要载体[18，35]。牛津 5K 的词汇被训练在巴黎的建筑物[30]，而对于“干净鞋子”词汇表上的数据集的随机子集教训。词汇量的大小是 20K，64 和 256 分别为 M1，M2 和 M3，按照相应的参考文献[40，19，18]。此外，我们还运行一个版本与密集采样 RGB-SIFT descrip-费舍尔矢量。

1The 性质与对应的所有者。所述图像示这里只用于科学目的。



(a)

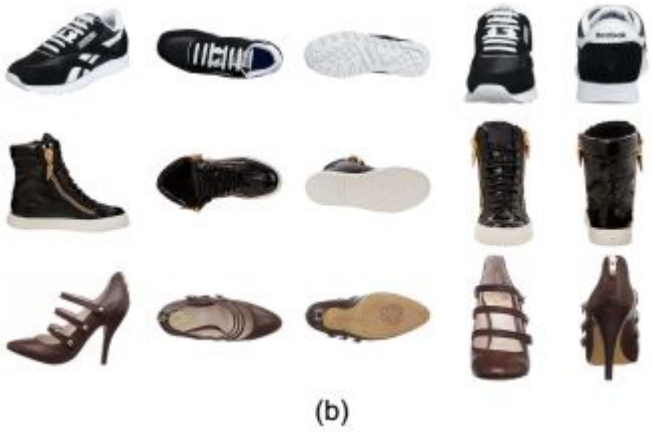


图 1：（一）从牛津 5K 两栋建筑的例子，和（b）从“干净鞋子”三款运动鞋的例子。那里存在于更广泛观点变异鞋的图像。

器[43]，费舍尔-D 表示。对于 M4，我们执行 CNN 在[22]提出并充分利用第二输出连接层作为图像表示。 CNN 的是使用 ImageNet 类培训。搜索性能正在使用中平均精度（MAP）测量。

结果与讨论。表 1 总结了结果在 Oxford5k 和“干净鞋子”。 ExpVLAD 采用与 20K 视觉单词和指数相似功能的大量词汇。其结果是，仅在靠近描述符对测量的两个相似的特征空间事例子。这导致比其它更好的性能在 Oxford5k 其中，密切和当地相关描述符对确实存在。然而，在鞋的图像，其中接近，局部描述符的真正的比赛很少存在由于大的变化的外观，ExpVLAD 达到最低性能。无论 Triemb 和 Fisher 获得建筑相当不错的成绩，但鞋子的结果是低的。这是再次引起的事实，局部描述符匹配是鞋上的图像，其中大视角不可靠差异存在。Triemb 优于费舍尔，在[19]的意见是一致的。在这项工作中，我们不考虑 RN 正常化[19]，

method↓	Oxford5k	CleanShoes
ExpVLAD	**76.54**	16.14
Triemb	61.64	25.06
Fisher	56.72	20.94
Fisher-D	53.62	36.27
Deep	45.50	**36.73**

表 1：不同实例搜索方法的性能：ExpVLAD [40]，Triemb [19]，费舍尔[18]和深[22，6]。对于费舍尔载体，我们考虑两个版本。费舍尔表示有兴趣点的版本，SIFT 描述符，和 Fisher-D 采用密集采样 RGBSIFT 描述符。上 Oxford5k 的结果是基于我们自己的实现，与文献报道一致在[40，19，5,6]。ExpVLAD 达到更好的性能比别人对 Oxford5k，但给出了“干净鞋子”最低的结果。另一方面，深获得最佳性能在“干净鞋子”，但比其他人对 Oxford5k 较低的结果。

因为它需要额外的训练数据来学习的投影矩阵，它不会不影响我们在这里做出的结论。费舍尔-D 作品利用颜色信息和密集采样点上“干净鞋子”比费舍尔更好。颜色是一种有用的线索区分不同的鞋子，和密集采样比较好不是鞋子不具备丰富的兴趣点探测器纹理图案。然而，费雪 D 不改善对费舍尔 Oxford5k。总体而言，在鞋的

性能比低得多上的建筑物。更有趣的是，ExpVLAD 实现性能优于其他人 Oxford5k，但给出了“干净鞋子”最低的结果。另一方面，深成于在“干净鞋子”最佳性能，但低于结果其他人 Oxford5k。我们的结论是没有任何现有的方法在两个建筑物运行良好，为 2D 的例子近平面实例搜索，和鞋子，作为三维的一例全视图实例搜索。

4.属性的泛型实例搜索

属性已在分类被证明是有利当训练实例充分覆盖在低级别特征的外观变化空间[10，45，2]。在我们的问题，只有 1 例可用的和仍存在一个宽范围的外观的变化。至于第二个问题，我们提高了纸张，可以我们使用属性来解决泛型实例搜索？在本节中，我们侧重于中已知的东西搜索是使用特定类别的属性相同的类。

在文献中，两种类型的属性已经研究，语义特征[23，2]和非语义特征[45，36]。获取语义属性要求相当多的人的努力，有时领域的专业知识，很难扩展到大量的属性。此外，手动拾取属性不能保证是歧视下的任务考虑[45]。另一方面，非语义属性不需要人类注释和有能力用于识别任务[45，36]进行优化。对于某些任务，像零次学习[2]和图像检索由文字属性查询[38]，有必要使用人可以理解的属性。然而，在实例中搜索鉴于 1 图像查询，其语义是不是真的必要。在这项工作中，我们使用非语义数据驱动属性。从一组训练实例提供某一类，我们的目标是自动学习列表特定类别的属性，并使用这些属性从同一类别上新实例进行实例搜索。

我们认为，学习类的特定属性三个标准。作为第一准则，需要的属性要能够使实例之间的区别。该第二个标准是，该属性必须共享中视觉上相似的训练情况。属性特定于一个训练实例是不太可能是描述性为除几个共同训练情况不明实例。和共享需要限制只在视觉上类似的培训情况作为视觉不同实例之间潜在的常见模式是不太可能上新实例来检测。第三个标准是，所学习属性之间冗余需要是低的。考虑到上述三个标准，我们采用现有的方法[45]这与我们的考虑合身。给定 n 个训练实例来自同一类别和瞄准对于 k 属性，方法学的实例属性映射

$$\underset{A}{\text{maximize}} \quad f_1(A) + \lambda f_2(A) + \gamma f_3(A), \tag{1}$$

where $f_1(A)$, $f_2(A)$ and $f_3(A)$ are defined as follows:

$$\begin{aligned} f_1(A) &= \sum_{i,j}^{n} \|A_{i\cdot} - A_{j\cdot}\|_2^2, \\ f_2(A) &= -\sum_{i,j}^{n} S_{ij} \|A_{i\cdot} - A_{j\cdot}\|_2^2, \\ f_3(A) &= -\|A^T A - I\|_F^2. \end{aligned} \tag{2}$$

i • 是第 i 个实例的属性表示。 F1（A）确保实例可分。骶髂关节表示实例 i 和实例ĵ之间的视觉相似，测量先验低层次的功能空间和 f2（一）鼓励类似属性外观相似的实例之间的交涉，诱导共享属性。 F3（A）处罚属性之间大幅裁员。获取实例属性映射，其中每一列代表一个属性后，K 属性分类的教训。博学的分类应用未知实例来实例生成属性表示，和搜索在做属性

空间。

数据集。我们评估的鞋子，汽车和建筑学会特定类别的属性。对于鞋，我们认为该“干净鞋子”前一节中所述。对于汽车，我们收集的 270 车 1110 图片来自易趣。我们记它由汽车。一辆汽车的一个图像被视为查询和我们的目标是找同车的其他图像。图 2 示出了两个 cars2 的一些例子。在建筑方面，我们通过收集所有的 567 图像组成一个小数据集从 Oxford5k 55 牛津地标。我们通过 OxfordPure 表示它。我们重用 Oxford5k 定义的 55 查询。对于训练鞋特有的属性，我们将收集的 300 鞋 2100 图片来自亚马逊，同一来源在哪里我们收集“干净鞋子”。对于学习专车专用属性，我们收集的 300 辆 1520 图片来自易趣。用于学习特定建筑物的属性，我们使用的一个子集大型建筑数据集介绍[6]。我们随机挑选每班 30 张图像，并自动选择 300 班这是最相关的 OxfordPure 根据视觉相似。我们总共 8756 图像，有些落得网址是坏了，有的班级只有不到 30 例。对于所有的鞋子，汽车和建筑物，在实例评估集合是不存在的训练集。

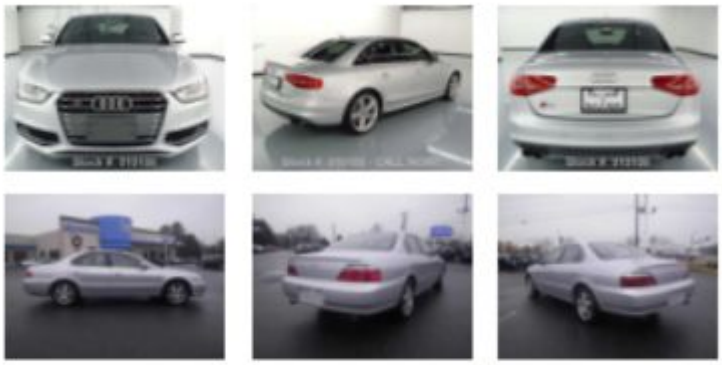


Figure 2: Examples of two cars.

实施细则。我们使用费舍尔矢量[35]与密集采样的 RGB-SIFT [43]作为底层表示学习属性分类。同样费舍尔表示用于选择用于学习特定建筑物的属性有关的训练示例。视觉在式（2）的相似性矩阵 S 被构建为相互 60-NN 邻接矩阵。两个训练之间的接近实例被计算为间的平均相似性在费希尔矢量空间中的两个实例的图像。

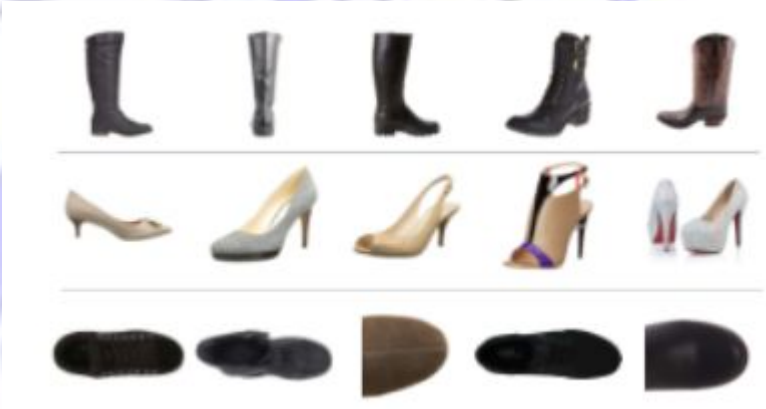
2The 性质与对应的所有者。所述图像示这里只用于科学目的。

attributes↓	*dim*	CleanShoes
Manual	40	18.99
Randomized	40	28.15
Learned	40	37.59
Randomized	1000	55.36
Learned	1000	56.56

表 2：不同属性的表现。的结果随机属性是 5 次运行的平均值。数据驱动非语义属性优于手动定义的属性。该了解到属性[45]达到最佳的性能，但是，如果属性的数量较高时，小在随机属性的性能差异允许跳过学习阶段。

结果与讨论。在第一个实验中，我们比较得知属性[45]两种选择，手动定义的属性和随机属性，上“干净鞋子”。对于手动定义的属性，我们使用由[14]中提出的属性列表。我们手动标注 2100 训练图像。在基准，42 的属性限定。但是，我们合并超高和高“上”和“脚跟高度”，因为它是难以注释超高和高为两个不同的属性，从而产生 40 属性。以产生随机化的属性，我们随机分成训练实例分为两组，假设在一组实例中有一个共同的视觉方面，其中其它实例没有。这种随机属性也已在[10]考虑用于图像分类。如表 2 所示，具有相同数量的属性，数据驱动的非语义属性的工作比手动属性显著更好。据悉属性比随机那些好得多当属性的数目是低的。随机拆分不走考虑到训练的基本视觉接近实例和属性不能在新的概括以及实例。这样的问题被减轻当大量的分裂被考虑。图 3 示出了三个学习属性。虽然属性没有明确的语义，这意味着，他们抓住鞋子之间的共同模式。

在第二个实验中，我们比较与“干净鞋子”，汽车和 OxfordPure 前一节中评估现有的方法中了解到的属性。表 3 示出了结果。属性表示工作比其他显著更好在鞋的数据集和汽车数据集。属性是在解决出众所造成的大摄像大出现变化角差存在于鞋和汽车的图像，甚至尽管属性是从其他实例得知。该图 3：三了解到属性。每一行都是一个属性和鞋子是具有高响应的人该属性。尽管自动学习属性没有任何语义，显然他们捕捉共享类似的鞋子之间的模式。第一属性表示
靴子。第二个属性描述了高跟鞋和第三个捕获圆头。属性表现也是行之有效的建筑物。此外，其他人相比，属性具有代表性低得多的维数。我们的结论是使用自动提出的方法学到特定类别的属性比其他方法更通用。



method↓	*dim*	CleanShoes	Cars	OxfordPure
ExpVLAD	—	16.14	23.70	**87.01**
Triemb	8064	25.06	18.56	75.33
Fisher	16384	20.94	18.37	70.81
Fisher-D	40960	36.27	20.89	67.41
Deep	4096	36.73	22.36	59.48
Attributes	1000	**56.56**	**51.11**	77.36

表 3：学会属性性能和现有的方法[40，19，18，22]。属性达到更好性能比其他鞋和汽车，并且看齐与其他建筑物。

五．.类别和泛型实例搜索属性

在本节中，我们考虑寻找一个实例从含有来自不同类别的实例的数据集。作为特定类别的属性进行了优化使同一类的实例之间的区别，他们可能无法分辨感兴趣的实例其他类别的实例。为了解决与实例查询实例的可能的混淆其他类别，我们还建议使用的类别级别信息。

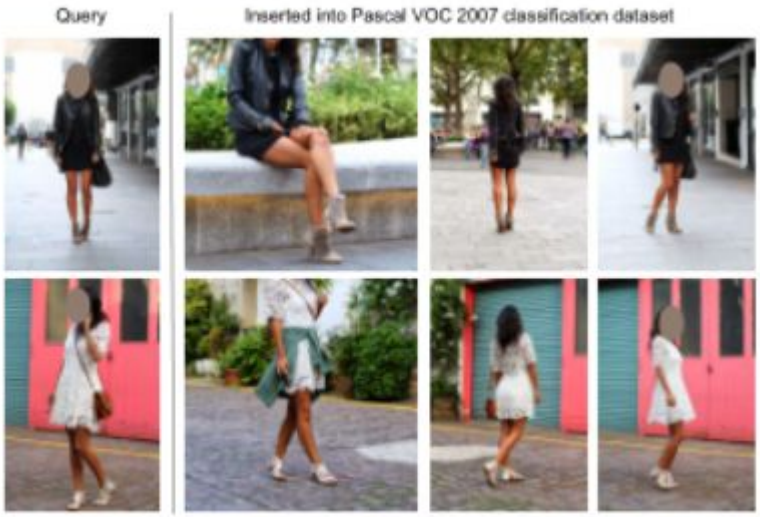


图 4：街头鞋子的两只鞋子的例子。左边被查询图像，其余插入帕斯卡尔 VOC 2007 分类数据集[9]提供牵引的图像。我们只考虑在查询鞋段例如，以确保目标是明确的。

我们考虑两种方式来捕获类别级信息。首先，我们采用的 4096 维输出查询插入到帕斯卡尔 VOC 2007 年分类数据集 CNN 的[22]作为附加的功能，因为它已经显示了的激活的第二完全连接层一个 CNN 的顶层捕捉高层次类别的相关信息[47]。 CNN 的使用 ImageNet 类培训。其次，我们建立了一个总类分类器缓解深学习特征的潜在问题，即深度学习功能可能带来的例子与查询实例，即使他们共同的要素是不相关的，如皮鞋和天空建筑物。与特定类别的属性相结合的两种类型的类别级信息，查询 q 和在搜索集的实施例 D 之间的相似性被计算

S（Q，D）= Sdeep（Q，D）+的 sclass

（D）+ SATTR（Q，D），（3）
其中 S 深（Q，D）为 q 和 d 在深的相似性学习特征空间的 sclass（d）是对 D 和 SATTR（Q，D）的分类响应于该属性的相似性空间。
数据集。我们评估鞋和建筑物。一个小的 15 鞋，总共 59 集的图像是由两个收集时尚 blogs3。这些图像被记录在街道上杂乱的背景，从“干净”的形象不同“干净鞋子”。我们认为鞋子的一个图像作为查询，目标是找到相同鞋等图像。鞋图像被插入到帕斯卡的 VOC 2007 分类数据集[9]的测试和验证部分。帕斯卡数据集提供牵引的图像。我们指含有鞋图像和分心作为 S 街头鞋子数据集。图 4 示出了两个例子。要了解鞋的分类，我们用 300'干净'鞋的属性在上一节为正的例子中学习，并考虑培训帕斯卡 VOC 2007 年的分类数据集作为反面教材的一部分。

在建筑方面，我们使用 Oxford5k。训练大楼分类，我们使用一节为正的例子中学习属性中的所有训练图像，并考虑将图像从帕斯卡 VOC 2007 年分类数据集作为反面典型。在建筑形象训练集不干净，还含有像天空元素和树木，我们预计建筑分类将不作为可用作鞋分类器。实施细则。我们只考虑对象在查询图像中区域，以确保目标是清楚的。它值得一提的是，虽然只是在部分对象查询图像被认为是，我们不能完全得到皮去掉一些鞋和天空的一些建筑物。我们使用选择性搜索[42]，产生许多候选在每个数据库中的图像位置，并在本地搜索图像作为[40]。我们采用具有 128 很短的表示尺寸。具体而言，我们减少的维度深学习的特点和属性交涉与 PCA 减少。与 Fisher 载体，我们采用在提出的美白方法[15]，证明优于 PCA。我们重用属性分类从以前部分。

结果与讨论。结果示于表 4.在街头的鞋子，特定类别的属性与两种类型的类级信息相结合的所提出的方法实现了最佳的性能，30.45 中
地图。我们观察到，仅仅考虑深特征时作为类别级信息，该系统带来了许多外观的例子。鞋分类培训了'干净'鞋的图像能够有效地消除这些不相关的前查询图像查询段前 5 个结果。

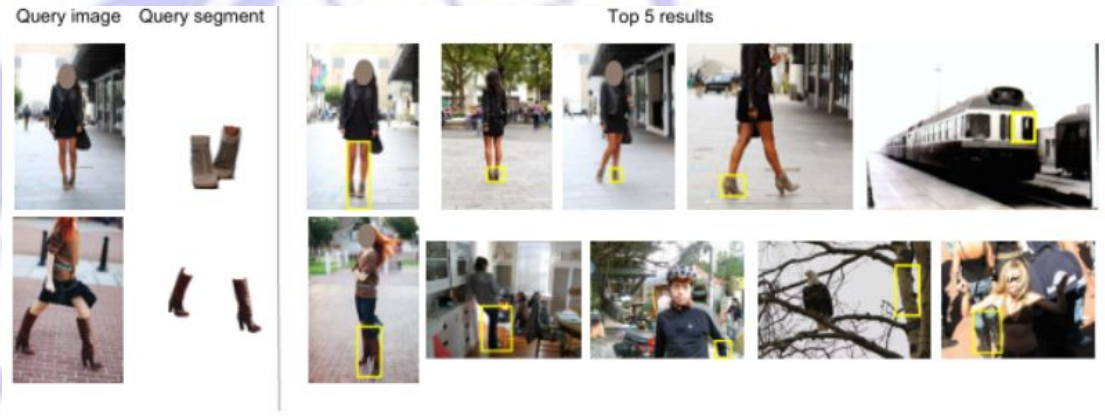


图 5 显示了前 5 名成绩通过该方法返回两个查询实例。在 Oxford5k 中，同样的方法实现了第二个最佳性能，65.14。

表现最好的是通过结合只有深深的功能类别的特定属性获得。建筑物分类并不能帮助可能是由于两个原因。一在该建筑的分类是在杂乱的图像学与天空和草，因此不能很好地处理深功能问题之前提及。如果我们有干净的建筑图像进行训练，分类会可能有助于提高性能。另一个原因是该图像在 Oxford5k 很大一部分是建筑而分类可能推不相关的建筑的例子到排名列表的顶部。该建筑将分级可能是在其中包含一个大的数据集的更多有用许多非建筑的例子。这 Oxford5k 包含许多建筑形象也是事实，解释了为什么指定类别单独的属性已经可以取得相当不错的表现，58.37。总体而言，我们得出结论，提出的具体类别的属性相结合方法两种类型的类别级信息是有效的，超越的类别级信息与组合.

References

[1] Large scale visual recognition challenge. http://www.
imagenet.org/challenges/LSVRC/2010, 2010.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Labelembedding for attribute-based classification. In CVPR, 2013.

[3] R. Arandjelovic and A. Zisserman. Multiple queries for large
scale specific object retrieval. In BMVC, 2012.

[4] R. Arandjelovic and A. Zisserman. Three things everyone ´
should know to improve object retrieval. In CVPR, 2012.

[5] R. Arandjelovic and A. Zisserman. All about VLAD. In ´
CVPR, 2013.

[6] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky.
Neural codes for image retrieval. In ECCV, 2014.

[7] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In ECCV,
2010.

[8] M. Douze, A. Ramisa, and C. Schmid. Combining attributes
and fisher vectors for efficient image retrieval. In CVPR,
2011.

[9] M. Everingham, L. Van Gool, C. K. I. Williams,
J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http:
//www.pascal-network.org/challenges/VOC/
voc2007/workshop/index.html.

[10] A. Farhadi, I. Endres, D. Hoiem,

and D. Forsyth. Describing
objects by their attributes. In CVPR, 2009.
[11] V. Ferrari and A. Zisserman. Learning visual attributes. In
NIPS, 2008.
[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic
segmentation. In CVPR, 2014.
[13] A. Gordoa, J. A. Rodr´ ıguez-Serrano, F. Perronnin, and
E. Valveny. Leveraging category-level labels for instancelevel image retrieval. In CVPR, 2012.
[14] J. Huang, S. Liu, J. Xing, T. Mei, and S. Yan. Circle &
search: Attribute-aware shoe retrieval. ACM Transactions on
Multimedia Computing, Communications, and Applications
(TOMM), 11(1):3, 2014.
[15] H. Jegou and O. Chum. Negative evidences and co- ´
occurences in image retrieval: The benefit of pca and whitening. In ECCV, 2012.
[16] H. Jegou, M. Douze, and C. Schmid. Hamming embedding ´
and weak geometric consistency for large scale image search.
In ECCV, 2008.
[17] H. Jegou, M. Douze, C. Schmid, and P. P ´ erez. Aggregating ´
local descriptors into a compact image representation. In
CVPR, 2010.
[18] H. Jegou, F. Perronnin, M. Douze, J. S ´ anchez, P. P ´ erez, and ´
C. Schmid. Aggregating local image descriptors into compact codes. TPAMI, 34(9):1704–1716, 2012.
[19] H. Jegou and A. Zisserman. Triangulation embedding and ´
democratic aggregation for image search. In CVPR, 2014.
[20] A. Joly and O. Buisson. Logo retrieval with a contrario visual
query expansion. In MM, 2009.
[21] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch:
Image search with relative attribute feedback. In CVPR,
2012.
[22] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS,
2012.
[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to
detect unseen object classes by between-class attribute transfer. In CVPR, 2009.
[24] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu,
L. Kennedy, A. Hauptmann, and J.

Curtis. Large-scale concept ontology for multimedia. MultiMedia, IEEE, 13(3):86–
91, 2006.

[25] P. Over, G. Awad, J. Fiscus, G. Sanders, and B. Shaw.
Trecvid 2012 - an introduction of the goals, tasks, data, evaluation mechanisms and metrics. In TRECVID, 2012.

[26] M. Perdoch, O. Chum, and J. Matas. Efficient representation
of local geometry for large scale object retrieval. In CVPR,
2009.

[27] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale ´
image retrieval with compressed fisher vectors. In CVPR,
2010.

[28] F. Perronnin, J. Sanchez, and T. Mensink. Improving the ´
fisher kernel for large-scale image classification. In ECCV,
2010.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In CVPR, 2007.

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman.
Lost in quantization: Improving particular object retrieval in
large scale image databases. In CVPR, 2008.

[31] D. Qin, S. Gammeter, L. Bossard, T. Quack, and
L. Van Gool. Hello neighbor: accurate object retrieval with
k-reciprocal nearest neighbors. In CVPR, 2011.

[32] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In ECCV,
2012.

[33] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. arXiv preprint arXiv:1403.6382, 2014.

[34] J. Revaud, M. Douze, and C. Schmid. Correlation-based
burstiness for logo retrieval. In MM, 2012.

[35] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Im- ´
age classification with the fisher vector: theory and practice.
IJCV, 105(3):222–245, 2013.

[36] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In ECCV, 2012.

[37] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Mobile product image
search by automatic query object extraction. In ECCV, 2012.

[38] B. Siddiquie, R. S. Feris, and L. S.

Davis. Image ranking and
retrieval based on multi-attribute queries. In CVPR, 2011.
[39] J. Sivic and A. Zisserman. Video google: A text retrieval
approach to object matching in videos. In ICCV, 2003.
[40] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders.
Locality in generic instance search from one example. In
CVPR, 2014.
185
[41] G. Tolias, Y. Avrithis, and H. Jegou. To aggregate or not ´
to aggregate: Selective match kernels for image search. In
ICCV, 2013.
[42] J. R. R. Uijlings, K. van de Sande, T. Gevers, and A. W. M.
Smeulders. Selective search for object recognition. IJCV,
104(2):154–171, 2013.
[43] K. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating
color descriptors for object and scene recognition. PAMI,
32(9):1582–1596, 2010.
[44] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han.
Contextual weighting for vocabulary tree based image retrieval. In ICCV, 2011.
[45] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang.
Designing category-level attributes for discriminative visual
recognition. In CVPR, 2013.
[46] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang. Weak
attributes for large-scale image retrieval. In CVPR, 2012.
[47] M. D. Zeiler and R. Fergus. Visualizing and understanding
convolutional networks. In ECCV, 2014.
[48] S. Zhang, M. Yang, X. Wang, Y. Lin, and Q. Tian. Semanticaware co-indexing for image retrieval. In ICCV, 2013.