Unsupervised Visual Representation Learning by Context Prediction

1. 摘要

使用图像空间信息作为一种自由且丰富的监督信号，来训练获得丰富的图像视觉表征。给定大的无标签的数据集，我们在每张图像上随机抽取成对的patch，来训练卷积神经网络，去预测第二个patch相对于第一个patch的位置。我们证明了，要在这个任务上表现的好，网络必须学习重建物体和他们的部分。使用这种参考图像内部信息的方法学习到的特征表示，可以捕捉到图像之间的视觉相似性。

1. 问题重述

基于监督学习的视觉表征学习：CV方法利用大量标注过的数据学习丰富高性能的视觉表征。然而标注太难。使用无监督方法无需标注，但是还表现不佳。毕竟没有标签信息，模型都不知道应该去学习什么。所以我们要研究无监督范式的视觉表征学习。

在文本领域，上下文（context）已经被证明在学习表征的时候，是一种非常强力的监督信号源。给出一大群文本信息，通过训练模型把一个单词映射到一个特征向量，这样就可以在给定向量的时候很容易的预测上下文中的单词。这个方法把一个明显的无监督问题（找到单词之间的好的相似度量），转到了一个“自我监督问题”：学习一个从给定的单词到它周围单词群的函数。这里上下文预测任务就仅仅是一种“pretext”来迫使模型学习一个好的单词嵌入。这种任务已经被证实了对许多真实的任务是有用处的，例如语义单词的相似性。

1. 解决思路

我们想要提供一个相似的“自我监督”的形式来处理图像数据：一个监督的任务，涉及为一个图像块预测他的上下文。我们从八个配置里面随机采样一个patch对，把他们展现给机器学习模型，不提供关于这对patch在图像中的原始位置信息。这样，算法必须去猜测他们之间的相对关系。我们的潜在假设是，在这个任务上很好的模型，需要理解场景和物体。也就是说，在这个任务上的好的视觉表征，需要去抽取图像中的物体和他们的部分，为了去推理他们的相对位置关系。我们借鉴文本领域的上下文学习，去解决视觉表征学习的问题。

1. 相关工作

认为一个好的图像表示法的一种方法是一个适当的生成模型的潜在变量。一个理想的自然图像生成模型既能根据图像的自然分布生成图像，又能简明扼要地寻求不同图像的共同原因并在它们之间共享信息。然而，即使是相对简单的模型，推断一个图像的潜在结构也是难以实现的。为了处理这些计算问题，许多工作，如唤醒-睡眠算法[25]、对比发散[24]、深度玻尔兹曼机[48]和变异贝叶斯方法[30，46]使用抽样来执行近似的推断。生成模型在较小的数据集，如手写数字上显示出良好的性能[25, 24, 48, 30, 46]，但没有一个被证明对高分辨率的自然图像有效。

无监督的表征学习也可以被描述成学习一个嵌入（即每张图像的特征向量），其中语义相似的图像是接近的，而语义不同的图像是相距甚远的。建立这样一个表征的方法是创建一个有监督的 "借口 "任务，使解决该任务的嵌入也对其他现实世界的任务有用。

例如，去噪自动编码器[56, 4]使用从噪声数据中重建作为借口任务：算法必须将图像与其他具有类似物体的图像连接起来，以区分噪声和信号之间的差异。稀疏自动编码器也使用重建作为借口任务，同时使用稀疏惩罚[42]，并且这种自动编码器可以被堆叠以形成深度表示[35, 34]。(然而，只有[34]成功地应用于全尺寸图像，仅发现三个物体就需要一百万个CPU小时）。我们认为，目前基于重建的算法在处理低层次的现象，如随机纹理方面很吃力，甚至很难衡量一个模型是否生成良好。

另外一种辅助任务是“上下文预测”。这种辅助任务存在于文本领域，‘[skip-gram](https://zhuanlan.zhihu.com/p/27234078)’模型展现出了很好的性能。他的思路是训练一个模型去从一个单一词汇去预测前面的n个词和后面的n个词。原则上，类似的推理可以应用于图像领域，这是一种视觉上的 "填空 "任务，但是，同样地，人们会遇到确定预测本身是否正确的问题[12]，除非人们只关心预测非常低层次的特征[14, 33, 53]。为了解决这个问题，[39]通过对一个图像区域的周边区域的反式近邻的共识投票来预测该区域的外观。我们以前的工作[12]明确地制定了一个统计测试，以确定数据是由预测还是由低层次的空假设模型更好地解释。

这些方法必须解决的关键问题是由于同一语义对象可能产生大量的像素，预测像素比预测单词要难得多。在文本领域，一个有趣的想法是将纯粹的预测任务转换为识别任务[41, 9]。在这种情况下，借口任务是将真实的文本片段与随机替换了一个词的相同片段区分开来。将其直接扩展到二维可能是区分真实的图像和一个补丁被数据集中其他地方的随机补丁所取代的图像。然而，这样的任务是微不足道的，因为区分低层次的颜色统计和照明就足够了。为了增加任务的难度和高度，在本文中，我们在同一图像中取样的多个可能配置的斑块之间进行分类，这意味着它们将共享照明和颜色统计。

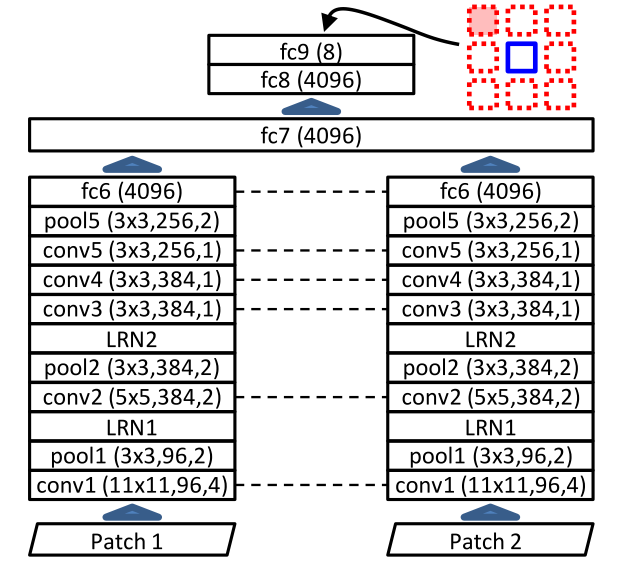
在无监督的图像学习方面的另一项工作是利用手工制作的特征和各种形式的聚类来发现物体类别（例如[51, 47]通过视觉词包学习生成模型）。这样的表述失去了形状信息，会很容易发现诸如叶子的集群。随后的一些工作试图使用与形状更密切相关的表征[36, 43]，但依赖于轮廓提取，这在复杂的图像中是困难的。许多其他方法[22, 29, 16]集中于定义相似性指标，这些指标可以用于更标准的聚类算法；例如，[45]将该问题重铸为频繁项集挖掘。地理学也可用于验证图像之间的联系[44, 6, 23]，尽管这对于可变形的物体来说可能会失败。

视频可以为表象学习提供另一种提示。对于大多数场景，即使外观随时间变化，物体的身份也不会改变。这种时间上的一致性在视觉学习文献中有着悠久的历史[18, 59]，同时代的工作显示了对现代检测数据集的有力改进[57]。

最后，我们的工作与一系列关于破坏性补丁挖掘的研究有关[13, 50, 28, 37, 52, 11]，该研究强调弱监督作为物体识别的一种手段。与目前的工作一样，他们强调在学习完整的物体和场景之前学习斑块（即物体部分）的表征的效用，并认为场景级别的标签可以作为一个前置任务。例如，[13]训练检测器对不同地理区域的敏感度，但实际目标是发现建筑风格的具体元素。

1. 详细阐述
   1. 学习视觉内容预测

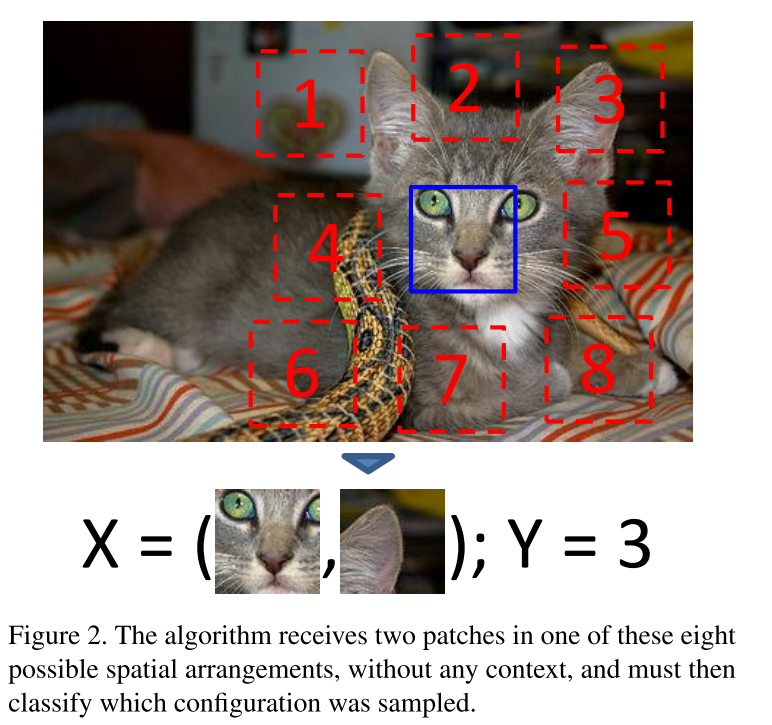
我们的目标是为我们的预文本任务学习一个图像表征，即预测图像中斑块的相对位置。文本任务，即预测图像中斑块的相对位置。我们采用卷积神经网络（ConvNets），众所周知，它能以最小的人为特征设计来学习复杂的图像表征。构建一个能够预测一对斑块的相对偏移的卷积网络，原则上是很简单的：网络必须将两个输入斑块通过几个卷积层，并产生一个输出，为可能被采样的八个空间配置（图2）中的每一个分配一个概率（即一个softmax输出）。然而，请注意，我们最终希望学习单个斑块的特征嵌入，这样，（在不同的图像中）非常相似的斑块在嵌入空间中会很接近。



为了获得给定图像的训练实例，我们在不参考图像内容的情况下，对第一个补丁进行统一采样。在不考虑图像内容的情况下，均匀地对第一个补丁进行采样。给定第一个补丁的位置，我们从八个可能的相邻位置中随机抽取第二个补丁，如图2所示。

* + 1. 避免细微的解决方案

在设计一个借口任务时，必须注意使确保任务迫使网络提取所需的信息（在我们的例子中，是高级语义），而不采取 "微不足道 "的捷径。在我们的案例中，低层次的线索，如边界图案或斑块之间的纹理，有可能成为这种捷径。因此，对于相对预测任务来说，在斑块之间加入一个间隙（在我们的例子中，大约是斑块宽度的一半）是很重要的。即使有间隙，相邻斑块之间的长线也有可能泄露正确答案。因此，我们也随机地将每个补丁的位置抖动7个像素（见图2）。

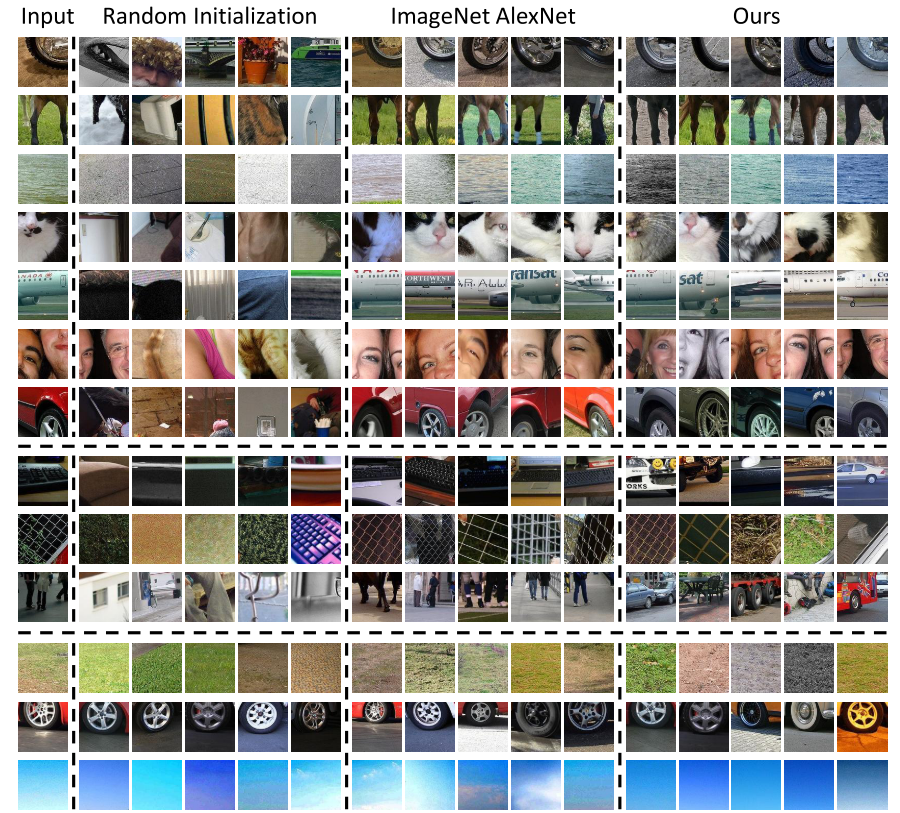


然而，即使是这些预防措施也是不够的：我们惊讶地发现，对于某些图像来说，存在着另一种特殊的解决方案。我们把问题追溯到一个意想不到的罪魁祸首：色差。色差产生于镜头对不同波长的光的聚焦方式的差异。在一些相机中，一个颜色通道（主要是绿色）相对于其他通道向图像中心收缩。事实证明，ConvNet可以通过检测绿色和洋红色（红+蓝）之间的分离度来学习相对于镜头本身对一个补丁的定位（见第4.2节）。一旦网络学会了镜头上的绝对位置，解决相对位置任务就变得微不足道了。为了处理这个问题，我们试验了两种类型的预处理方法。一种是将绿色和品红色向灰色转移（"投影"）。具体来说，让a = [-1, 2, -1]（RGB空间中的 "绿-玛瑙色轴"）。然后我们定义B = I - aTa/(aaT)，它是一个矩阵，可以减去一种颜色在绿-玛瑙色轴上的投影。我们将每个像素值乘以B（Hermit矩阵变换）。另一种方法是在每个补丁中随机丢掉3个颜色通道中的2个（"丢色"），用高斯噪声（标准偏差为剩余通道标准偏差的1/100）取代丢掉的颜色。对于定性结果，我们展示了 "丢色 "的方法，但发现两者的表现相似；对于物体检测结果，我们展示了两种结果。

**实施细节**：ImageNet数据集，仅使用数据，不使用标签。首先，我们将每张图片的大小调整到15万到45万总像素之间，保留长宽比。从这些图像中，我们对分辨率为96×96的斑块进行采样。为了提高计算效率，我们只对网格状的斑块进行采样，这样每个采样的斑块可以参与多达8个独立的配对。我们允许网格中的采样斑块之间有48个像素的间隙，但也会使网格中每个斑块的位置在每个方向上抖动-7至7像素。我们通过(1)平均减法(2)投射或删除颜色（见上文）来预处理斑块，以及(3)随机降低一些斑块的采样，使其总像素低至100，然后再向上采样，以建立对像素化的稳定性。当应用简单的SGD来训练网络时，我们发现网络的预测会退化为对8个类别的统一预测，fc6和fc7的所有激活都变成了0。这意味着优化会完全卡在一个鞍点上，它忽略了来自低层的输入（这有助于最小化最终输出的方差），因此，网络无法调整低层的特征并逃离这个鞍点。因此，我们的最终实现采用了批量归一化[26]，没有scale和shift（γ和β），这迫使网络的激活在不同的例子中变化。我们还发现，高动量值（如.999）加速了学习。在实验中，我们使用一个在K40 GPU上训练了大约四周的ConvNet。

1. 实验
   1. 最近邻

回顾我们的直觉，训练应该把相似的表征分配给语义相似的斑块。在本节中，我们的目标是了解我们的网络认为哪些斑块是相似的。我们从随机抽取96x96个斑块开始，我们用fc6的特征来表示这些斑块（即我们重新移动图3中所示的fc7和更高的斑块，并且只使用两个堆栈中的一个）。我们使用这些特征的归一化相关来寻找最近的邻居。一些斑块（从1000个随机查询中选出）的结果显示在图4中。为了比较，我们使用在ImageNet上训练的AlexNet的fc7特征（通过对斑块进行加样得到），以及使用我们架构中的fc6特征，但没有进行任何训练（随机权重初始化），重复了这个实验。如图4所示，我们的特征所返回的匹配结果往往能捕捉到我们所追求的语义信息，在语义内容上与AlexNet相匹配（在某些情况下，例如汽车车轮，我们的匹配结果能更好地捕捉到姿势）。有趣的是，在少数情况下，随机（未经训练的）ConvNet也表现得相当好。



图表 4 通过最近的邻居获得的补丁集群的例子。查询的补丁显示在最左边。匹配的是三种不同的特征：来自我们架构的随机初始化的fc6特征，在标记的ImageNet上训练后的AlexNet fc7，以及从我们的方法中学到的fc6特征。查询是从1000个随机抽样的斑块中选择的。最上面的一组是我们的算法表现良好的例子；对于中间的AlexNet胜过我们的方法；而对于最下面的三个特征都表现良好。

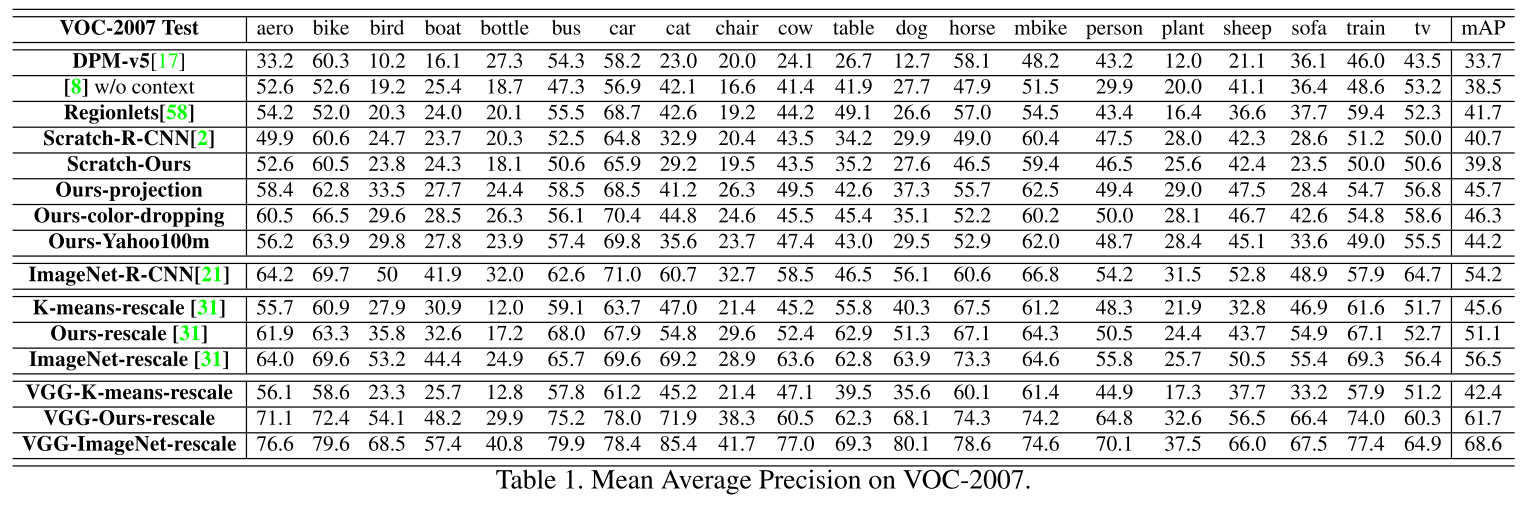
* 1. 旁观者清：色差的可学习性。

我们在早期的近邻实验中注意到，有些斑块检索到的是来自图像中同一绝对位置的匹配斑块，不管内容如何，因为这些斑块显示出类似的畸变。为了进一步证明这一现象，我们训练了一个网络来预先确定从ImageNet中取样的斑块的绝对（x，y）坐标。虽然这个回归器的总体准确度不是很高，但它对某些图像的表现却令人惊讶：对于前10%的图像，平均（均方根）误差为0.255，而偶然表现（总是预测图像中心）产生的RMSE为0.371。图5显示了一个这样的结果。应用建议的 "投影 "方案使前10%的图像的误差增加到0.321。

* 1. 目标检测

之前关于Pascal VOC挑战的工作[15]表明，在ImageNet上进行预训练（即训练ConvNet来解决ImageNet的挑战），然后对网络进行 "微调"（即针对PASCAL数据重新训练ImageNet模型），比单独在Pascal训练集上的训练有很大的提升[21, 2]。然而，据我们所知，没有任何作品表明，无论使用多少数据，对图像进行无监督的预训练都能提供这样的性能提升。

表1显示了我们的结果。我们的架构从头开始训练（随机初始化）的表现比从头开始训练的AlexNet略差。然而，我们的预训练弥补了这一点，将从头开始的数字提高了6%MAP，并且比在Pascal上从头开始训练的AlexNet式模型高出5%以上。这使我们比用ImageNet标签预训练的R-CNN的性能落后8%[21]。这是我们所知道的在VOC 2007上没有使用数据集以外的标签的最好结果。我们运行了其他的基线，用批处理规范化初始化，但发现它们的表现比图中的差。



在上述微调实验中，我们通过估计conv-和fc-层的均值和方差，然后重新调整权重和偏置，使conv和fc层的输出对每个通道的均值为0，方差为1，从而去掉了批量归一化层。然而，最近的工作[31]根据经验表明，在微调之前，权重的缩放会对测试时的性能产生很大的影响，并认为我们以前去除批量归一化的方法导致权重的缩放太差。他们提出了一种简单的方法，在不改变网络计算功能的情况下重新调整网络的权重，这样网络在微调期间表现得更好。使用这种技术的结果显示在表1中。他们的方法对所有的方法都有促进作用，但对已经经过良好扩展的ImageNet类别模型的促进作用较小。请注意，在这个比较中，我们使用了fast-RCNN[20]以节省计算时间，并且我们的模型中丢弃了所有预训练的fc层，用[31]的K-means程序重新初始化它们（该程序被用来初始化 "K-means-rescale "行中的所有层）。因此，在微调和测试期间，所有模型的网络结构都是一样的。

考虑到我们基本上有无限的数据来训练我们的模型，我们可能期望我们的算法也能为更高容量的模型，如VGG[49]提供很大的提升。为了测试这一点，我们按照[49]的16层结构训练了一个模型，用于网络每一侧的卷积层（最后的fc6-fc9层与图3相同）。我们再次使用fast-rcnn对Pascal VOC上的表示进行了微调，只转移了卷积层，并再次按照Kr¨ahenb¨uhl等人[31]的方法对转移的权重进行了重新调整，并初始化了其余部分。作为基线，我们用ImageNet预训练的16层模型[49]进行了类似的实验（尽管我们保留了预训练的fc层，而不是重新初始化它们），同时也用K-means[31]初始化了整个网络。请注意，用K-means初始化的模型与类似的AlexNet模型表现大致相同，这表明大部分的提升来自于无监督的预训练。

* 1. 几何学估算

第4.3节的结果表明，我们的表征对物体是敏感的，尽管它最初并没有被训练来寻找这些物体。这就提出了一个问题。我们的表征是否能提取对其他非物体的任务有用的信息？为了找出答案，我们对我们的网络进行了微调，以执行Fouhey等人[19]提出的对NYUv2的表面法线估计，遵循Wang等人[57]的微调程序（因此，我们直接与那里报告的无监督预训练结果进行比较）。我们使用了色彩退化网络，如第4.3节所述，重组了全连接层。令人惊讶的是，我们的结果几乎等同于使用完全标记的ImageNet模型得到的结果。一个可能的解释是，ImageNet分类任务对鼓励网络关注几何形状的作用相对较小，因为一旦物体被识别，几何形状就基本不重要了。这方面的进一步证据可以在图4的第七行看到：ImageNet AlexNet的近邻都是汽车车轮，但它们与查询的补丁并不一致。

* 1. 视觉数据挖掘

视觉数据挖掘[44, 13, 50, 45]，或无监督对象发现[51, 47, 22]，旨在使用大型图像集合来发现恰好描述相同语义对象的图像片段。其应用包括数据集可视化、基于内容的检索，以及需要将视觉数据与其他非结构化信息（如GPS坐标[13]）联系起来的任务。对于自动数据挖掘，我们在第4.1节中的方法是不够的：尽管物体斑块与相似的物体相匹配，但纹理与相似的纹理也同样容易匹配。然而，假设我们从同一个物体上抽取了两个不重叠的斑块。不仅两个斑块的近邻列表会共享许多图像，而且在这些图像中，近邻会处于大致相同的空间配置。另一方面，对于纹理区域来说，邻居的空间配置是随机的，因为纹理没有全局布局。

为了实现这一点，我们首先从图像中抽取一个由四个相邻斑块组成的星座。我们首先从图像中抽取四个相邻的斑块（我们用四个斑块来减少偶然发生的匹配空间排列的可能性）。我们找到前100张与所有四个斑块最匹配的图像，忽略空间布局。然后，我们使用一种几何验证方法[7]来过滤掉四个匹配的图像在几何上不一致的地方。因为我们的特征在语义上更有针对性，所以我们可以使用比[7]弱得多的几何验证类型。最后，我们通过计算前100个匹配的几何验证次数来对不同的星座进行排名。

1. 评析