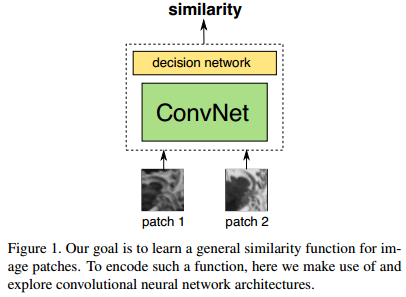
CVPR 2015 被引：230

Abstract

在这篇论文中我们展示了如何直接地从图像数据（比如，无需借助于手动设计地特征）中学习一般的相似性函数（similarity function）来用于比较图像块，这对于许多的计算机视觉问题是一个至关重要的任务（a task of fundamental importance）。为了编码这样的函数，我们选择了一个基于CNN的模型，该模型经过训练可以应对（account for）图像外观的广泛地变化。为此，我们探索和研究了多种神经网络的架构，这些架构专门适用于此任务。我们表明，这种方法可以在几个问题和基准数据集上表现明显优于最先进的技术。

**Introduction**

比较图像之间的块大概是最基本的任务（one of the most fundamental tasks）之一在计算机视觉和图像分析中。它经常被用作是一类子程序（subruntine），在一个广泛类别的视觉任务中扮演着重要的角色。这些任务的范围从（range from）低水平的任务如SFM，宽基线匹配（wide baseline matching），建筑全景图，和图像超分辨率到更高级的任务例如对象识别，图像检索，和对象类别分类，提到几个有特点的例子。

 当然，决定两个块是否对应的问题是相当有挑战性的因为存在太多的因素影像一个图像最终的外观。这些因素包括视点的改变，场景整体照明的变化，遮挡（occlusions），阴影（shading），相机设置的不同，等等。事实上，这种比较块的需要已经导致（given）了许多手工设计的特征描述符的发展在过去的年份中，包括SIFT，它在计算机视觉社区有巨大的影响。然而，这种手动设计的描述符也许不能以最佳的方式顾及（take into account）上述所有因素。另一方面，当下，人们可以容易地访问（甚至使用可用软件生成）包含图像间块对应关系的大型数据集。这引出了（beg）下面的问题：我们可以合理的使用这些数据集来自动地学习图像块地相似函数吗？

这篇论文的目标是肯定地（affirmatively）解决上面地问题。我们的目标因此是能够产生一个块的相似性函数从头（from scratch），比如，不企图用任何手动设计的特征而是直接学习这个函数从带注释（annotated）的原始图像块的对中。为此，也受到神经架构和深度学习最近进展的启发，我们选择用深度卷积神经网络来表示这样的函数（如图1）。在这样的过程中，我们也感兴趣解决这样一个问题，在像这样的任务中，什么样的网络架构最适合使用。我们因此探索和提出了不同的网络类型，其架构具有不同的权衡和优势（trade-offs and advantages）。在所有的情况下，为了训练这些网络，我们使用包含原始图像块对的（匹配和不匹配）大型数据集作为单独的输入。这允许更进一步地提高我们方法的表现简单地通过用更多地例子丰富这个数据集（因为用于自动生成这样的样本的软件是随时可以得到的）。

总结这部分，文章的主要贡献如下：（1）我们直接地从图像数据（即，没有任何手动设计的特征）中学习一个针对块的通用的（general）的相似性函数，它可以隐含的考虑各种类型的变换和影响（由于，比如，宽基线，照明等等）。（2）我们探索和提出了各种各样的不同的神经网络模型适用于表示这样的函数，同时突出了提供改进性能的网络架构，如[19Learning local feature descriptors using convex optimisation]。（3）我们将我们的方法应用在不同问题和基准数据集（benchmark datasets），表明它明显（significantly）优于最先进的方法，并且它产生（leads to）的特征描述符有更好的表现比起手动设计的描述符（比如SIFT，DAISY）或者其他学习得到得描述符就像在[19]中。重要地是，由于它们的卷积本质（nature），得到的描述符即使在一个密集方式下也能非常有效地计算。

**2 Related work**

比较块的传统的方法是用描述符和平方欧式距离。大多数特征描述符是手工制作的像SIFT或者DAISY。最近，学习描述符的方法已经被提出来[27 2012]（比如，类似DAISY的描述符学习池化区域和降维）。Simonyan等人[19]提出了一个凸的程序来训练两者。

我们的方法，然而，是受最近成功的卷积神经网络的启发.。尽管这些模型在训练期间涉及高度非凸的目标函数，它们还是在不同的任务中表现出很好（outstanding）的结果。Fischer等人[2014]分析了来自AlexNet网络卷积描述符的表现（它是在Imagenet数据集上训练的）在众所周知的Mikolajczyk数据集上，并且显示这些卷积描述符在大多数情况下除了模糊都比SIFT表现得要好。它们也提出了一个非监督训练方法来获取描述符，它表现得比SIFT和Imagenet训练网络都好。

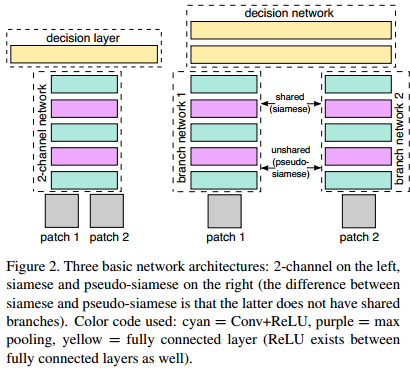
Zbontar和LeCun在[28 MC-CNN-1]中最近提出了一种基于CNN的方法来比较块为了计算在小基线（small baseline）立体问题中的代价并且在KITTI数据集中展现了最好的表现。然而，那个工作的专注点仅仅是比较由非常小的块组成的对，比如在窄基线（narrow baseline）立体中的那些块。相反，这里我们针对相似函数，即那些能够应付更广的外观改变和被用在更广泛和更具挑战性的应用中，包括，例如，宽基线立体[[1]](#footnote-1)，特征匹配和图像检索。

**3.Architectures**

正如已经提到的，神经网络的输入被认为是一对图形块。我们的方法没有设置任何限制关于通道的数量在输入的块上，即给定具有颜色补丁的数据集，可以训练网络以进一步提高性能。然而，为了能够比较我们的方法和在现存数据集上最新的方法，我们仅仅选择灰度级的块在训练期间。而且，除了在3.2节中描述的SPP模型，在所有其他的情况下，块给到作为网络的输入，被假定有固定的大小即64×64（这意味着原始的块也许需要重新调整大小以适应上述空间维度）。

有许多方式块对可以被网络所处理以及在这种情况下如何进行信息共享。出于这个原因，我们探索和测试了不同的模型。我们开始于章节3.1通过描述三种基本的神经网络架构，我们研究的，即2-channel，Siamese，Pseudo-siamese（见图2），它们提供了不同的权衡（trade-off）在速度和精度方面（注意，通常情况下，被应用的块匹配（patch-matching）技术暗示了测试一个块对比（against）大量的其他块，所以重复使用计算信息通常是有用的）。本质上这些架构产生于它们中的每一个试图解决下面问题的不同方式：当组成（composing）一个相似函数来比较图像块，我们该首先选择计算一个描述符为每一个块然后在这些描述符的上面创建相似性还是我们或许选择跳过描述符计算的部分直接进行相似性估计？

除了上面提到的基本模型，我们也

在3.2节中描述了一些额外变化和网络架构相关的。这些变化，彼此不是相互排斥的，能够与3.1节中描述的基本模型结合使用。总的来说，这

导致了不同的模型，可能被用于比较图像块的任务。

**3.1 Basic models**

**Siamese：**这种类型的网络效仿（resembles）了有描述符的思想。网络中有两个分支，共享相同的结构和相同的权重集。每一个分支将两个块中一个作为输入然后应用一系列的卷积，ReLU和最大池化层。分支的输出被连接（concatenated）并被给到一个顶网络（a top network）由线性全连接和ReLU层组成的。在我们的测试中，我们使用了一个顶网络由两个线性全连接层（每一个有512个隐藏单元）组成，它俩被一个ReLU激活层分隔开。

Siamese网络的分支们能够被看作描述符计算模块而顶网络被看作相似函数。对于在测试时匹配两组块的任务，描述符能够被首先独立地计算使用分支，然后用顶网络匹配（甚至是用一个距离函数像l2[[2]](#footnote-2)）。

**Pseudo-siamese：**就复杂性而言，这种架构能够被认为在siamese和2-channel网络之间。更具体地说，它有上面描述地siamese网络的结构除了两个分支的权重是分开的，即不共享。这增加了能够在训练时被调整的参数的数量，并且比起一个受限制的的siamese网络提供了更多的灵活性，但不如接下来介绍的2-channel网络多。另一方面，它保持了在测试时的siamese网络的效率。

**2-channel:** 不像前面的模型，这里没有直接的描述符概念在架构里。我们简单地将一对输入的两块看作一个双通道的图像，它被直接地馈送到网络的第一个卷积层。在这种情况下，网络的底部由一系列的卷积，ReLU和最大池化层组成。这个部分的输出然后被作为顶部模块的输入，它仅仅由一个带着一个输出的全连接线性决定层组成。这个网络提供了更好的灵活性比上面的模型因为它从共同处理两个补丁开始。而且，它训练得很快，但总体上在测试时它更昂贵（expensive）因为它需要所有块的组合相互之间测试以一种粗暴的方式（in a brute-force manner）。

**3.2 Additional models**

**深度网络（Deep network）。**我们采用了由Simonyan和Zisserman在[20 2014]提出的技术，它建议将更d的卷积层打破成为3×3的核，由ReLU激活分隔，它被认为提高了网络内部的非线性性并且使决定函数（decision function）更加有判别力。他们也报道了初始化这样的一个网络也许是困难的，我们，然而，没有观察到这种现象并且也没有按照惯例从头开始训练网络。在我们的案例中，我们对我们的模型应用了这种技术，最后的结构的卷积部分是（turns out to）由一个卷积的4×4层和6个拥有着3×3层的卷积层组成，被ReLU激活层隔开。我们也会稍后再实验结果中看到，网络结构中的这样一个改变能够更进一步的提高表现，这与在[20 2014]中做的类似的观察一致。

**中心环绕双流网络（Central-surround two-stream network）**.正如它的名字所展示的那样，提出的架构包含两个分开的流，中心和周围，这使得能够在两个不同分辨率的空间区域进行处理。更进一步来说，中心高分辨率流接收两个32×32的块作为输入，这些块是通过裁剪（在原始分辨率上）每个输入的64×64的块的中心的32×32的块生成的。此外，环绕的低分辨的流接收两个32×32的块作为输入，它们是通过对原始输入块对的一半进行下采样生成的。得到的两个流能够被处理通过使用任何在3.1节中提到的基础架构。（看图3作为一个例子，为每个流使用了一个siamese架构）。

一个原因来使用这样的一个双流机构是因为多分辨率信息被认为是重要的在提高图像匹配的表现上。而且，通过两次考虑一个块的中心部分（即在高分辨率和低分辨率流）我们隐含地把更多地注意力靠近块地中心的像素上，更少地关注在外围的像素，这也能够有助于提高匹配的精确性（本质上，因为池化被用于下采样图像，外围的像素在匹配的过程中被允许有更大的变化）。注意，总的输入维度减少了两倍在这个例子中。结果，训练进行得更快，是另一个实际的优势。

**空间金字塔池化网用于计算块（SPP network for comparing patches.）**。到目前为止，我们一直假设网络需要有64×64固定大小的输入块。这个需求来自这样的实时，即网络的最后一个？上一个（the last convolutional layer？）卷积层的输出需要有一个预先定义的维度。因此，当我们需要比较任意大小的补丁时，这意味着我们首先需要重新调整它们到上述空间维度。然而，如果我们看看像SIFT似的描述符的例子，例如，我们能够发现另一个处理任意尺寸大块的方法是通过调整【19：34】空间池化区域的大小和输入块的大小成比例，这样我们依旧可以保持对最后一个卷积层的需要的固定的输出维度而并不会降低输入块的分辨率。

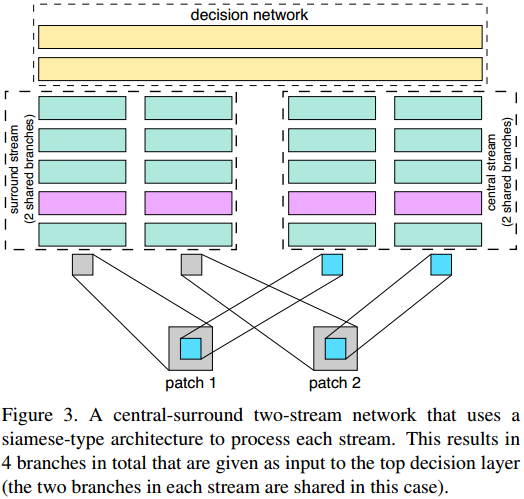
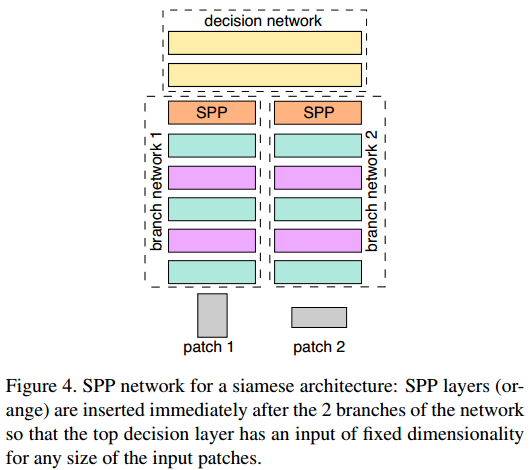
这也是最近提出的SPP网络结构[11 2014]的背后的思想，这本质上相当于在网络的卷积层和全连接层之间插入一个空间金字塔池化层。这样的一层聚集了最后一个卷积层的特征通过空间池化，这里面池化区域的大小是依赖于输入的大小的。受这个思想启发，我们建议也考虑根据上述SPP架构调整3.1节的网络模型。对于所有考虑的模型，这都很容易实现。（比如，看图4作为一个用siamese模型的例子）。【8：51 11 】=》【tomorrow 12：00-2：00】

**4. Learning**

**Optimization.**我们训练所有的模型以一种强有力的监督方式。我们使用了一个基于hinge的损失术语和平方L2范式正则化，能够得到下面的学习目标函数

,

其中，w是神经网络的权重，是关于第i个训练样本的网络输出，是对应的标签（-1和1分别表示不匹配和匹配的对）。

 ASGD（异步随机梯度下降）有固定的学习率1，动量0.9，权重衰减λ=0.0005被用来训练模型。训练以128大小的小批量进行。权重被随机初始化所有的模型都从头开始训练。

**Data Augmentation and preprocessing** 为了克服过拟合，我们扩充了训练数据通过水平和垂直翻转两个补丁并旋转到90度，180度，270度。因为我们没有观察到过拟合当以这样的方式训练时，我们以一个确定的迭代次数训练我们模型，一般是两天，然后在测试集上测试表现。

训练集的大小允许我们直接在GPU内存中存贮所有的图片并在训练过程中非常有效地检索块对。图像被及时？程序运行过程中（on-the fly）扩充。我们在Torch中使用Titan GPU，卷积例行程序（routines）取自Nvidia cuDNN库。我们的siamese描述符在GPU上只比在CPU上计算SIFT描述符慢两倍并根据[10]比起在GPU上的Imagenet描述符快两倍。

**5.Experiments**

我们应用我们的模型到各种各样的模型和数据集。在下面我们报告结果并提供了和最先近技术的比较。【201904092127】=》【201904101419】

**5.1. Local image patches benchmark**

对于我们模型的第一个分析，我们使用了标准基准数据集（standard benchmark dataset）来自[3:2010]，它由三个子集组成，Yosemite，Notre Dame，和Liberty，它们中的每一个包含超过450000个图像对（64×64）在高斯差分特征点周围取样。块被尺度和方向标准化（normalized）。每个子集被生成通过使用实际的3D对应被获取通过多视图立体深度图。这些图被用来产生500000个真值特征点对为每个数据集，拥有相等数量的正（正确的）和负（错误的）匹配。

为了评价我们的模型，我们使用了[4]的评价协议并生成ROC曲线通过给距离设置阈值在描述符空间的特征对之间。我们报道了在95%召回率时的假正率对6个训练和测试集的组合中的每一个，以及所有组合的平均值。我们也报告了均值，被记作mean（1，4），仅用于[1],[3]中使用的4种组合。（在这种情况下，训练只在Yosemite或者Notre Dame上进行，不在Liberty上）。

表1公布了几个模型的表现，也详述了他们的架构（我们也实验了更小的核，更少的池化层，和添加正则化，并没有注意到在表现上的显著提高）。我们简要地总结一下可以从表中得出的结论。第一个重要的结论是基于2-channel的架构（比如，2ch，2ch-deep，2ch-2stream）清楚得展示了最好的表现在所有模型上。这表明从网络的第一层直接联合使用两个块的信息是重要的。

2ch-2stram网络是表现最好的网络在这个数据集上，2ch-deep紧随其后（这证明了在匹配过程中多分辨率信息的重要性和提高网络深度也有帮助？？）。事实上，2ch-2stream的表现地比之前最新的技术好一大截，比[19：co：2014]高出2.45倍。和SIFT的差别甚至更大，在这个例子中我们的模型给出了6.65倍好的得分。（SIFT得分在mean（1，4）上是31.2根据[3]）。

关于基于siamese的架构，这些也能实现更好的表现比起已有的最好的系统。这是相当有意思的因为，比如这些siamese网络没有尝试学习池化区域的形状、大小和位置（像，例如[19,3]做的那样），而是仅仅使用标准的最大池化层。在这些siamese网络中，二流网络（siam-2stream）有最好的表现，再次证实了在比较图像块时多分辨率信息的重要性。而且，psudo-siamese网络（pseudo-siam）比对应的siamese（siam）更好。【201904101430】

我们也进行了另外的实验，在里面我们测试了siamese模型的表现当它们的顶层决定层被由网络的两个分支产生的两个卷积描述符的l2欧式距离所替换（在名字中以后缀l2标记）。在这种情况下，在应用欧几里得距离之前，描述符是被l2范式的（我们也测试了l1范式的）。对于pseudo-siamese,只有一个分支被用来抽取描述符。就像预料的那样，在这种情

况中双流网络（siam-2stream-l2）计算了更好的距离比起samese网络（siam-l2），而后者，计算了更好的距离比起pseudo-siamese模型（pseudo-siam-l2）。事实上，siam-2stream-l2网络甚至比之前最先进的描述符[19: co]表现得更好，这相当得惊人考虑到这些siamese模型从没有用L2距离训练过。

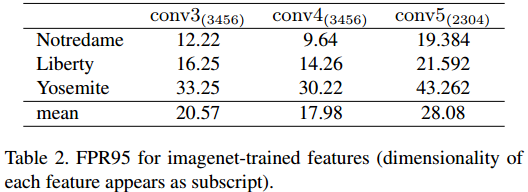
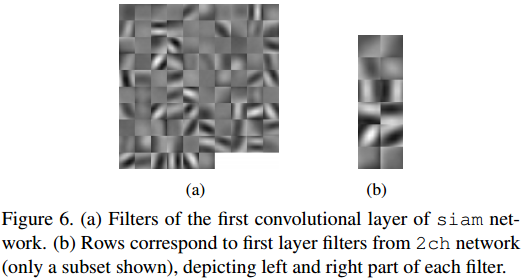
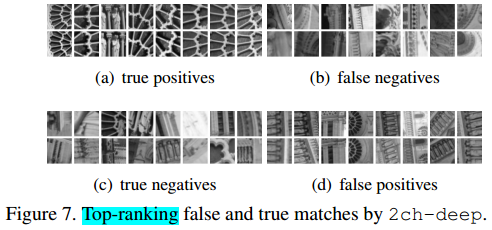
 对于不同模型的一个更细节的比较，我们提供了对应的ROC曲线在图5中。而且，我们在表2中展示了imagenet训练的CNN特征（这些被l2标准化来提高效果）。 在这些之中，conv4给出了最好的FPR95分数，等于17.98。这比SIFT要好，但仍比我们的模型差很多。

图6（a）展示了被siamese网络学到的第一个卷积层的过滤器（filters）。而且，图6（b）展示了左边和右边部分为两通道网络2ch的第一层过滤器的一个子集。值得提的是对应的左右部分看起来彼此不同（negative to each other）,这从根本上说意味着网络已经学到了计算特征的差别在两个块之间（但注意，并非所有的2ch的第一层过滤器都像这样）。最后，我们在图7中展示了一些顶级的（top ranking？）的错误的和正确的匹配当用2ch-deep网络计算的时候。我们观察到错误的匹配项可以被很容易地搞错甚至是被人类（注意，例如，在错误地正例中地两个块看起来是多么相似啊）



对于实验剩下的部分，我们指出我们使用了在Liberty数据集上使用的模型。

**5.2.Wide baseline stereo evaluation**

对于这个评估我们选择了Strecha et al[23]的数据集, 它包含了数个图像序列包含真值单应性（ground truth homographies）和激光扫描的深度图。我们使用了“fountain”和“herzjesu”序列来分别产生6个和5个校正的立体对。两个序列中的基线我们选择的都在增长使得匹配更加困难。我们的目标是展示用神经网络计算的光度成本与最先进的手工制作的特征描述符产生的成本相比是有优势的，因此我们选择和DAISY进行比较。

因为我们关注的不是效率，我们使用一个未优化的管线来计算光度代价。更具体地说，对于2-channel网络，我们使用了一个暴力方法，我们提取块在对应的极线上用亚像素估计，构建批（包含来自左图和来自右图对应极限上的所有块）并计算网络的输出，产生如下代价：

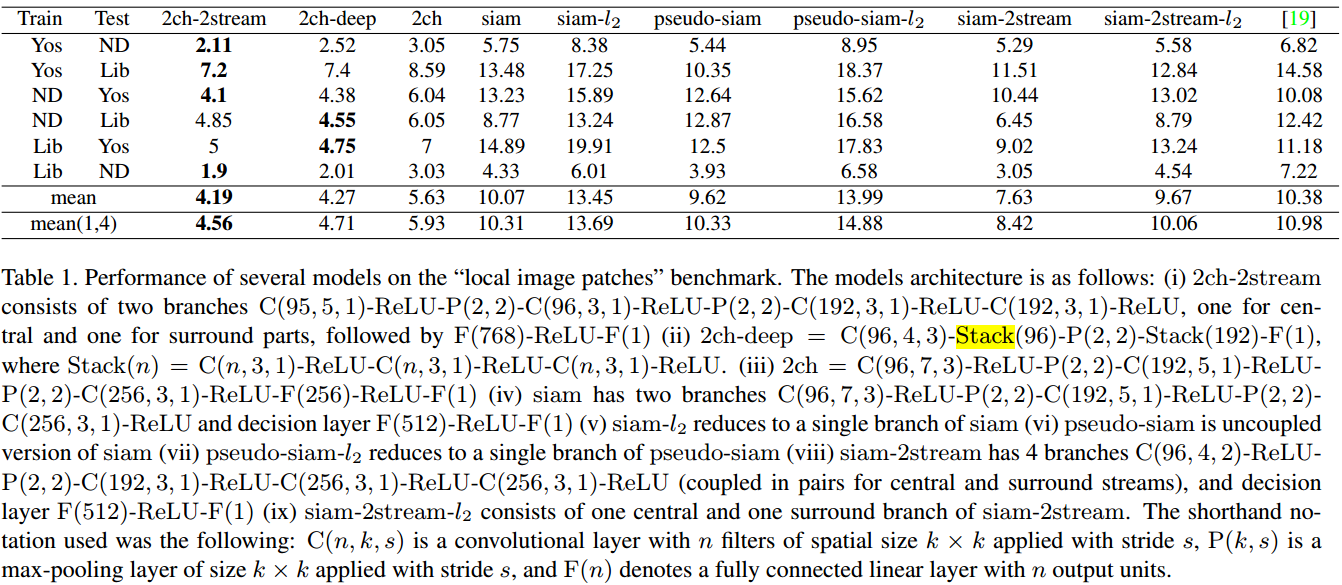
(2)

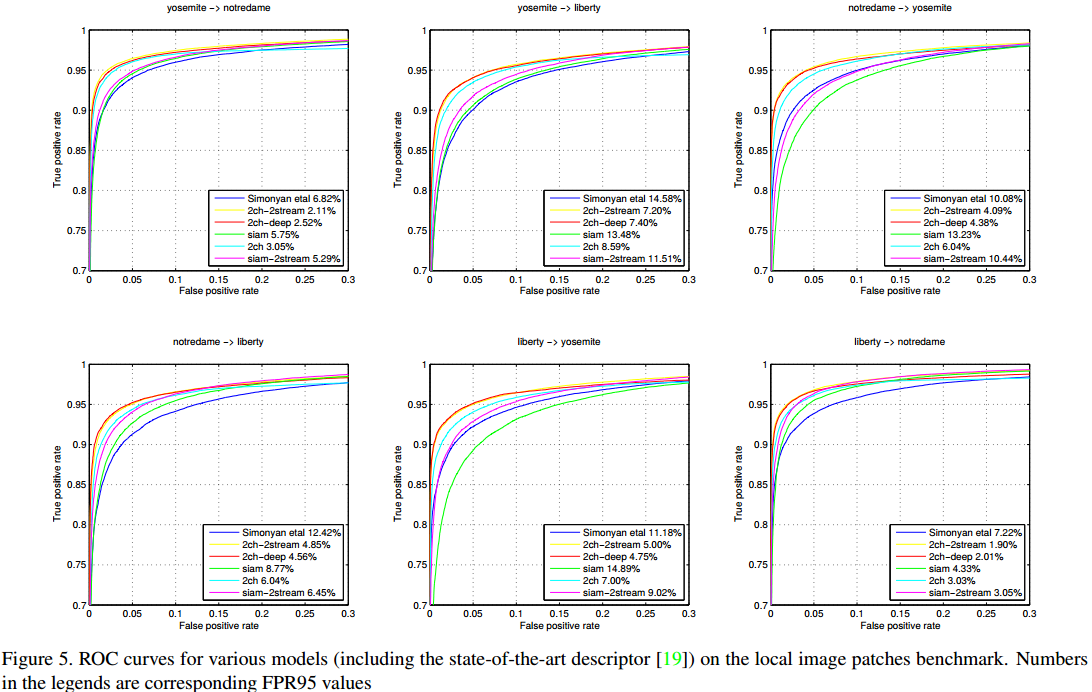
这里，I(P)指代邻域强度矩阵在一个像素p周围，是神经网络的输出给定一对块和，d是极线上点之间的距离。

对于siames类型的网络，我们计算了描述符对于每个像素在两幅图中一次，然后匹配他们通过决定层或者l2距离。在第一种情况下，光度成本的公式如下：

(3)

其中是顶层决定层的输出，和是siamese或者pseudo-siamese分支的输出，即描





述符（在siamese网络的情况下）。对于L2匹配，它变成（holds）：

(4)

值得注意的是上面所有的代价能够被计算得更有效通过使用加速优化类似于[28:mc-cnn]。这本质上意味着训练所有的全连接层作为1×1的卷积，计算siamese网络的分支仅仅一次，然后进一步计算这些分支的输出和网络最后的输出在所有位置，在全图像上使用多个前向传递。（即，对于2-channel架构，这样的一个计算光度代价的方法将仅仅需要喂给网络尺寸等于输入图像对的s2·dmax全双通道网络，其中，s是网络第一层的跨度，dmax是最大视差）。

一旦计算出，光度成本随后在被用作一元项在下面成对的MRF能量中通过使用基于FastPD[12]的算法[8]最小化。）（我们设置，，σ = 7，Ԑ是一个四联通的网格 ）。

，

我们在图9中展示了一些关于计算的深度图定性的结果（有和没有全局优化）对‘fountain’数据集（“herzjesu”的结果出现在supp.材料中因为缺少空间）。全局MRF优化的结果直观地（visually）验证了光度代价通过神经网络计算的比起手工制作的特征更加鲁棒性，和由2-channel架构产生的深度图的高质量。没有全局优化的结果也显示估计的深度图包含更多的细节比起DAISY。它们也许呈现了一个非常稀疏的错误集合对于基于siamese网络的情况，但是这些错误能够被很容易地消除在全局优化的过程中。

图8也展示了定量的比较，在这种情况下集中在基于siamese的模型上因为它们更加有效（efficient）。该图的第一个图展示了（对于单个立体对）与真值的偏差分布在所有的误差阈值范围内（在这里表示为场景深度范围的分数？）。而且，这个图的其他图总结了相应的误差分布为6个增加基线的立体对。（在这种情况下我们也单独地展示了误差分布当只有非遮挡像素被考虑）。误差阈值被设置为3个和5个像素在这些图中。（注意到最大地视差在500像素左右在最长的基线上）。正如被看到的那样，所有的siamese模型表现得比DAISY好得多在所有的误差阈值上和所有的基线距离上（比如，注意对应图的曲线的不同）。

**5.3. Local descriptors performance evaluation**

我们也测试了我们的网络在Mikolajczjk数据集上来进行局部描述符评估[16:2005].这个数据集包含48个图像在6个序列里包括相机视点的改变，模糊，压缩，光照改变和缩放伴随着转变的逐渐增加。有已知的真值单应性在第一张和其他图像之间在序列里。

测试技术和[16]中的一样。简单说，为了测试一对图像，探测器（detectors）被应用到两张图片上来提取关键点。按照[10:2014]中，我们使用了MSET探测器。探测器提供的椭圆用于从输入图像中提取块。椭圆的尺寸被放大了3倍来包含更多的内容。然后，根据网络的类型，两个描述符，即siamese或者pseudo-siamese分支的输出，被提取出来，或者所有的块对被给到一个2-channel网络来分配分数。

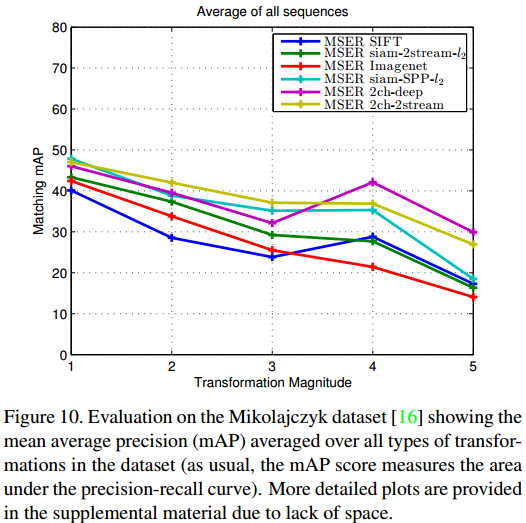
一个定量的比较在这个数据集上被展示对不同的模型在图10中。这里我们也测试了CNN网络siam-SPP-l2,它是一个基于SPP的siamese架构（注意siam-SPP和siam是一样的但是又额外的两个SPP层-见图4）。我们使用一个插入的SPP层，它的空间维度是4×4。正如被看到的那样，这提供了一个很大的提高在匹配的表现上，表明了在比较图像块时这种架构的巨大效用。关于剩下的模型，图10中的观察结果重新确认了前面实验已经得出的结论。我们只是再次注意到siam-2stream-l2的表现，它（尽管没用用l2距离来训练）能够显著地优于SITF的表现并且也能够跟上用imagenet-trained的特征的表现。（但是，使用的维度要低得多，是512）。

**6.Conclusion**

在这篇论文中我们展示了如何直接地从原始图像像素中学习一个通用地相似函数为块，它能够以一个CNN模型的形式编码。为了那个目的，我们学习了多个专门适用于这项任务的神经网络架构，并表明它们表现出极好的表型，显著优于最新的技术在不同的问题和基准数据集上。

在这些架构中，我们注意到2-channel-based的在结果中显然是优越的。因此，值得研究如何在未来进一步加快对这些网络的评估。关于基于siamese的架构，2-stream多分辨率模型非常强大，总是带来表现上的显著提高也证实在比较块时多分辨率信息的重要性。相同的结论被应用在基于SPP的siamese网络中，它也不断提高结果的质量。[[3]](#footnote-3)

最后，我们应该注意到，简单地使用一个更大的训练集就有可能够有利于并且更进一步地提高我们方法的总体表现（因为训练集在当前的实验中使用的确实被认为太小了以今天的标准）。



不变理论：物体识别的易用性不受观察者视点（viewpoint）的影响。

依赖理论：假设物体识别更快并且更容易当物体被看到从一个确定的角度。

不变视点机制（Invariant viewpoint mechanism）：当对象识别涉及容易的分类区分（例如汽车和自行车之间）时使用。

依赖理论机制 （Dependent viewpoint meanism）：当图像时复杂的和苦难的（diffidult）。

1. <https://www.cnblogs.com/meadow-glog/p/4331531.html>

   <http://www.robots.ox.ac.uk/~vgg/presentations/vanguard/18.9.1997/pcp/talk.html#p6>

   文末补充 [↑](#footnote-ref-1)
2. L2范式，即欧式距离 [↑](#footnote-ref-2)
3. 事实上，SPP的表现能够进一步地提高，因为没有宽高比补丁被使用在SPP模型训练期间。（这样地补丁仅仅在测试时出现） [↑](#footnote-ref-3)