

## 通过卷积神经网络学习比较图像斑块

谢尔盖·扎戈鲁科 (Sergey Zagoruyko)  
巴黎理工大学巴黎理工大学  
sergey.zagoruyko@imagine.enpc.fr

尼科斯·科莫达基斯  
巴黎理工大学巴黎理工大学  
nikos.komodakis@enpc.fr

### 抽象

在本文中，我们展示了如何直接从图像中学习数据（即，不求助于手动设计的功能）用于比较图像斑块的通用相似度函数，对于许多公司来说，这是一项至关重要的任务。推杆视觉问题。为了编码这样的功能，我们选择基于CNN的模型，该模型经过训练可以说明图像外观变化多样。为此，我们探索和研究多种神经网络架构，特别适合此任务的。我们证明这种方法可以大大胜过状态有关几个问题和基准数据集的最新技术。

### 1.简介

比较图像之间的补丁可能是其中之一计算机视觉和图像分析的最基本任务。它通常用作子程序，起着重要的作用。在各种各样的视觉任务中扮演的角色。这些范围可以从低级任务，例如运动结构，宽基线匹配，建立全景图和图像超分辨率，完成更高级别的任务，例如对象识别，图像提及对象类别的检索和分类一些典型的例子。

当然，决定两个补丁是否正确的问题是彼此之间是否相互反应是非常具有挑战性的。太多的因素影响了产品的最终外观。

图片 [17]。这些可能包括观点的改变，差异数量的整体照度，颜色，阴影，相机设置的差异等。实际上，这需要

补丁的比较已经引起了。在过去的几年中，许多手工设计的功能描述符，包括SIFT [15]，这对公司产生了巨大影响。普特视觉社区。但是，这种手动设计的脚本编写者可能无法考虑最佳确定上述所有因素的方式。补丁的外观。另一方面，如今可以轻松访问（甚至使用可用资源生成

可从以下网站在线获得源代码和经过训练的模型：  
<http://imagine.enpc.fr/~zagoruyk/deepcompare.html>（工作由EC项目FP7-ICT-611145 ROBOSPECT支持）。

### 相似

决策网络

### 卷积网

补丁1 补丁2

图1.我们的目标是学习im-年龄补丁。为了对这样的函数进行编码，我们在这里使用和探索卷积神经网络架构。

软件）包含补丁对应关系的大型数据集图像之间[22]。这提出了以下问题：可以我们适当地使用这些数据集来自动学习图像补丁的相似性功能？

本文的目的是肯定地解决以上问题。因此，我们的目标是能够产生一个从零开始修补相似性函数，即不尝试-可以使用任何手动设计的功能，而是从带注释的原始图像对中正确学习此功能补丁。为此，还受到了最近的研究进展的启发神经网络和深度学习，我们选择代表用深层卷积神经网络发送了这样的函数网络[4, 3]（图1）。这样做，我们也很感兴趣解决什么网络架构问题最好用于这样的任务。因此，我们探索并提出各种类型的网络，具有展示不同的取舍和优势。在任何情况下，都要训练这些网络，我们使用大型数据库作为唯一输入包含成对的原始图像补丁（均匹配和不匹配）。这样可以进一步提高仅通过丰富此数据库即可实现我们方法的有效性包含更多样本（作为自动生成的软件，这种样品很容易获得[21]）。

总结本节，论文的主要贡献如下：（i）我们直接从图像数据中学习（即，没有任何手动设计的功能）

---

## 第2页

可以隐含考虑到补丁的补丁功能  
计算各种类型的转换和效果（由于  
例如较宽的基线，照度等）。(ii) 我们探索  
并提出各种不同的神经网络模型  
适应于代表这样的功能，突出显示  
同时提供改进的  
形式。如[19]。(iii) 我们在以下几个方面采用我们的方法：  
常规问题和基准数据集，表明它  
远远超过了最新技术  
以比具有更好的性能的描述符为特色  
手动设计的描述符（例如SIFT，DAISY）或其他  
学习的描述符，如[19]。重要的是，由于他们的  
由于具有滚动性质，因此生成的描述符非常有效  
以密集的方式进行计算。

### 2.相关工作

比较补丁的常规方法是使用  
描述符和平方欧氏距离。最具特色  
描述符是按SIFT手工制作的[15]或DAISY [26]。  
最近，学习描述符的方法已经被提倡  
构成[27]（例如，类似DAISY的描述符学习重新合并  
gion和降维[3]）。Simonyan等。  
[19]提出了在这两个任务上进行训练的凸过程。  
但是，我们的方法受到最近成功的启发  
卷积神经网络[18, 25, 24, 9]。虽然  
这些模型涉及高度非凸的目标函数-  
训练期间，他们在  
各种任务[18岁]。Fischer等。[10]分析了性能-  
来自AlexNet网络的卷积描述符  
（在Imagenet数据集[13]）  
已知的Mikolajczyk数据集[16]，并表明这些  
在大多数情况下，进化描述符的表现优于SIFT  
除非模糊。他们还提出了无监督的培训计划，  
得出优于两个SIFT的描述符的方法  
和Imagenet训练有素的网络。

Zbontar和LeCun在[28]最近提出了一个  
基于CNN的方法来比较补丁以进行计算  
小基线立体声问题的成本，并显示出最佳效果  
KITTI数据集中的性能。但是，重点是  
工作只是在比较由很小的对组成的对  
像窄基线立体声那样的音色。相反，  
在这里，我们的目标是可以解决的相似性函数  
外观变化范围更广，可用于  
范围更广，更具挑战性的应用程序集  
包括例如宽基线立体声，功能匹配和  
图像检索。

### 3.架构

如前所述，神经网络的输入  
被认为是一对图像补丁。我们的模型  
不要对数量施加任何限制  
输入补丁中通道的数量，即给定的数据集具有

可以训练网络的色块，以进一步  
提升性能。但是，为了能够比较我们的  
现有数据集上的最新方法  
我们选择在训练期间仅使用灰度色块。毛皮-  
Thermore，除了在  
第3.2节，在所有其他情况下，作为输入提供的补丁  
假定网络的固定大小为64×64  
（这意味着可能需要将原始补丁调整为  
以上空间尺寸）。

补丁对可以通过多种方式进行  
网络访问以及如何进行信息共享  
在这种情况下发生。因此，我们探索了  
测试了多种模型。我们开始在部分3.1由去  
刻画了三种基本的神经网络架构  
我们研究了2通道，暹罗语，伪暹罗语（请参阅  
图2），这在速度方面提供了不同的权衡  
和准确性（请注意，通常，应用补丁匹配  
技术暗示针对大量补丁测试补丁  
其他补丁，因此重新使用计算所得的信息  
有用的方法）。本质上，这些架构源自  
他们每个人尝试解决以下问题的不同方式  
较低的问题：在为  
比较图像补丁，我们首先选择计算  
每个补丁的描述符，然后在  
这些描述符的顶部还是我们可能选择跳过  
与描述符计算直接相关的部分  
进行相似度估算？

除了上述基本模型之外，我们还描述了  
在3.2节中，有关网络的一些额外变化  
建筑。这些变体不是相互存在的，  
彼此兼容，可以与任何  
3.1节中描述的基本模型。总的来说  
导致可以用于各种模型  
比较图像补丁的任务。

#### 3.1.基本型号

暹罗语：这种类型的网络类似于  
有一个描述符[2, 6]。网络中有两个分支机构-  
具有完全相同的架构和相同的作品  
一组权重。每个分支将两个中的一个作为输入  
补丁，然后应用一系列卷积ReLU  
和最大池化层。分支输出是串联的  
并提供给包含线性完全连接的顶级网络  
选定的层和ReLU层。在我们的测试中，我们使用了顶级网络  
由2个线性完全连接层组成（每个层有512个  
隐藏单元）由ReLU激活层分隔。

暹罗网络的分支机构可以看作是  
脚本计算模块和顶级网络-作为  
相似度函数。对于匹配两组的任务  
在测试时修补程序，描述符可以首先独立计算  
认真使用分支，然后与顶部匹配  
网络（甚至具有1 2之类的距离函数）。

伪暹罗语：就复杂性而言，这个架构

第3页

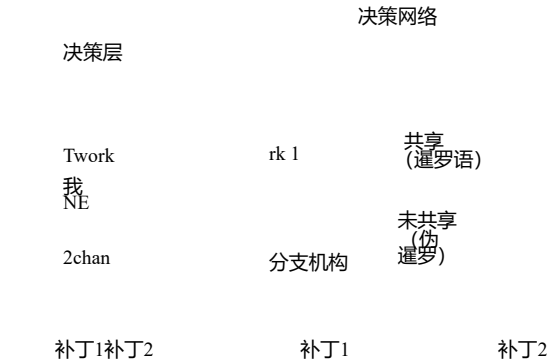


图2.三种基本的网络架构：左侧为2通道，暹罗语和伪暹罗语（之间的区别暹罗语和伪暹罗语是后者没有共享分支机构）。使用的颜色代码：青色= Conv + ReLU，紫色=最大合并，黄色=完全连接的层（ReLU存在于完全连接的图层）。

可以认为是暹罗之间和2通道网络。更具体地说，它具有上述暹罗网的结构，不同之处在于两个分支的权重是解耦的，即共享。这增加了可以在培训期间进行调整，并提供更大的灵活性比受限制的暹罗网络要多，但不如接下来介绍2通道网络。另一方面，它在测试时保持暹罗网络的效率。

2通道：与以前的型号不同，这里没有架构中描述符的直接概念。我们只是将输入对的两个补丁视为2通道图像，直接输入到第一卷积层网络。在这种情况下，网络的底部工作由一系列卷积，ReLU和max-池层。然后将这部分的输出作为输入到仅包含完全连接的顶部模块具有1个输出的线性决策层。该网络提供与上述模型相比，它具有更大的灵活性通过共同处理两个补丁。而且是训练速度很快，但通常在测试时更贵因为它要求对补丁的所有组合进行测试彼此以蛮横的方式。

3.2. 附加型号

深度网络。我们采用了Si-monian和Zisserman在[20]建议分手更大卷积层成较小的3x3内核，由ReLU激活，应该会增加非网络内部的线性关系并做出决策更具歧视性。他们还报告说可能是很难初始化这样的网络，但是，我们不这样做观察这种行为并从头开始训练网络照常。就我们而言，将这种技术应用于



图3.中央环绕的两流网络使用暹罗式架构来处理每个流。这导致总共4个分支被提供给顶级决策层（在这种情况下，每个流中的两个分支是共享的）。

模型，最终架构的卷积部分转向包含一层卷积4x4层和6张卷积具有3x3层的常规层，由ReLU激活层分隔位置。正如我们稍后将在实验结果中看到的那样，网络体系结构的这种变化可以有助于进一步提高性能，这符合在[20]。

中央环绕的两流网络。作为它的名字建议，建议的架构包含两个独立的中央和环绕声流，可用于处理发生在两个不同分辨率上的空间域位置。更具体地说，中央高分辨率流接收作为输入的两个32x32色块通过裁剪（以原始分辨率）中央32x32每个输入64x64色块的一部分。此外，环绕低分辨率流接收两个32x32色块作为输入，是通过对原始图像进行一半下采样而生成的对输入补丁。然后可以得到两个流通过使用所描述的任何基本体系结构进行处理在3.1节中（有关使用暹罗的示例，请参见图3）每个流的体系结构）。

使用这种两流体系结构的原因之一-事实是因为已知多分辨率信息是在提高图像匹配性能方面很重要-ing。此外，通过考虑补丁的中心部分两次（即在高分辨率和低分辨率下流），我们暗中将更多的注意力放在像素上到补丁的中心，而较少关注PE-中的像素外围，这也有助于提高精度匹配（本质上，由于将池应用于向下采样图像，允许外围像素具有匹配期间的差异更大）。注意总输入在这种情况下，二维尺寸减少了两倍。如结果，培训进行得更快，这也是彼此的



图4. 暹罗体系结构的SPP网络：SPP层（或-  
ange）插入到网络的两个分支之后  
这样顶层决策层就可以输入固定维数  
适用于任何大小的输入色块。

实际优势。

用于比较的空间金字塔池（SPP）网络  
补丁。到目前为止，我们一直假设  
网络要求输入补丁的大小固定  
64×64。此要求来自以下事实：  
网络需求的最后一个卷积层的输出  
具有预定的尺寸。因此，当我们  
需要比较任意大小的补丁，这意味着  
我们首先必须将它们调整为上述空间尺寸。  
但是，如果我们看一下SIFT这样的描述符的例子，  
例如，我们可以看到另一种可能的交易方式  
具有任意大小的补丁是通过调整  
与空间大小成正比的空间汇集区域  
输入补丁，以便我们仍然可以维护所需的  
最后一个卷积层的固定输出维数  
而不会降低输入色块的分辨率。

这也是最近提出的SPP-  
网络架构[11]，实际上相当于插入  
卷积之间的空间金字塔池化层  
层和网络的完全连接层。这样的  
层汇总了最后一个卷积层的特征  
通过空间池化，池化的大小重新  
gions取决于输入的大小。受此启发，  
我们建议还考虑调整以下网络模型：  
根据上述SPP架构的第3.1节。这个  
对于所有考虑的模型（例如，  
有关暹罗模型的示例，请参见图4。

#### 4.学习

优化。我们训练所有型号的超级超级  
可见的方式。我们使用基于铰链的损失项并平方  
12-范数正则化导致以下学习

目标函数

$$\frac{\lambda}{2} \|w\|^2 + \sum_{i=1}^n \max(0, 1 - y_i o_i), \quad (1)$$

其中w是神经网络的权重，邻网  
网络输出的第i个训练样本，和y<sub>i</sub>∈  
{-1, 1}对应的标签（-1和1表示a  
不匹配和匹配对）。

持续学习率1.0，动量0.9的ASGD  
权重衰减λ= 0.0005用于训练模型。  
训练以大小为128的迷你批次进行。  
随机初始化，并且从头开始训练所有模型。

数据扩充和预处理。  
蝙蝠过度拟合我们通过翻转两个  
成对的水平和垂直补丁并旋转到  
90、180、270度。因为我们不会注意到过度拟合  
以这种方式进行训练我们会训练一定数量的模型  
迭代，通常需要2天，然后测试性能  
对测试集进行操作。

训练数据集的大小使我们可以将所有图像存储在  
正确地放置在GPU内存中并非常有效地检索补丁  
训练期间进行配对。图像“实时”增强。  
我们在Torch中使用Titan GPU [7]和卷积例程  
摘自Nvidia cuDNN库[5]。我们的暹罗  
GPU上的脚本编写器仅比计算速度慢2倍  
CPU上的SIFT描述符，比Imagenet快2倍  
根据[10]。

#### 5.实验

我们将模型应用于各种问题，  
数据集。在下文中，我们将报告结果，并  
与最新技术进行比较。

##### 5.1. 本地映像补丁程序基准

为了首次评估我们的模型，我们使用了标准  
来自[3]由三个子集组成，  
优胜美地，巴黎圣母院和自由，其中每个包含  
采样了超过450,000个图像补丁（64 x 64像素）  
围绕高斯特征点的差异。补丁  
将比例尺和方向标准化。每个子集  
是使用获得的实际3D对应关系生成的  
通过多视图立体深度图。这些地图用来  
为每个数据集产生500,000个地面真实特征对，  
正数（正确）和负数（包含  
rect）匹配。

为了评估我们的模型，我们使用评估程序  
[4]并通过阈值生成ROC曲线  
描述符空间中特征对之间的距离。我们  
报告每个人的95%召回率（FPR95）的假阳性率  
训练和测试集的六个组合中的一个，以及  
所有组合的均值。我们还报告平均值  
表示为均值（1，4），仅适用于这4种组合

表1报告了几种模型的性能，以及还详细介绍了他们的架构（我们也进行了实验内核更小，最大池化层更少以及在不注意任何重大影响的情况下添加归一化性能证明）。我们简要总结一下可以从该表得出的结论。第一重要结论是基于2通道的架构（例如2ch，2ch深，2ch-2stream）清晰显示所有型号中最佳的性能。这是什么表明联合使用信息很重要来自网络第一层的两个补丁。

2ch-2stream网络是性能最高的网络在此数据集上，紧随其后的是2ch深（此版本证明了在处理过程中多分辨率信息的重要性匹配，这也有助于增加网络深度）。实际上，2ch-2stream的表现胜过之前的最先进的技术，达到2.45倍比[19]！与SIFT的差异甚至是更大，我们的模型在这方面的得分高出6.65倍案例（根据[3]）。

关于基于暹罗语的体系结构比现有的状态更好的性能最先进的系统。这很有趣，因为例如这些暹罗网络都没有试图学习形状，合并区域的大小或位置（例如[19, 3]），但仅使用标准的最大池布局-ers。在暹罗机型中，两流网络（siam-2stream）表现最佳，验证再次说明多分辨率信息的重要性当比较图像补丁时。此外，伪暹罗网络（pseudo-siam）更好比相应的暹罗一（siam）。

我们还进行了其他实验，其中我们测试了暹罗模型的性能顶层决策层被l2欧几里得dis-取代产生的两个卷积描述符的距离网络的两个分支（以suf-修复名称中的l2）。在这种情况下，在应用欧氏距离，描述符是l2-归一化的（我们还测试了l1标准化）。对于伪暹罗只分支用于提取描述符。如预期的那样这种情况下为两流网络（siam-2stream-l2）计算出的距离比暹罗网络更好（siam-l2），它反过来计算的距离比伪暹罗模型（pseudo-siam-l2）。事实上，siam-2stream-l2网络设法跑赢大市甚至以前的最新描述符[19]，即鉴于这些暹罗模型从未被训练使用l2距离。

要更详细地比较各种模型，

自由	16.25	14.26	21.592
优胜美地	33.25	30.22	43.262
意思	20.57	17.98	28.08

表2. FPR95用于图像网络训练的特征（维数每个功能都显示为下标）。

我们在图5中提供了相应的ROC曲线。毛皮-  
Thermore，我们在表2中显示了imagenet-的性能受过训练的CNN功能（将这些l2归一化以改善结果）。其中，conv4的FPR95得分最高，等于17.98。这使其比SIFT更好，但比我们的模型还差很多

(一种) (b)  
图6. (a) 暹罗网络第一卷积层的过滤器工作。(b) 行对应于2ch网络的第一层过滤器（仅显示了一个子集），描绘了每个过滤器的左右部分。

(a) 正面肯定 (b) 假阴性

(c) 真实否定词 (d) 误报

图7. 2ch-deep排名最高的错误和真实匹配。


图6 (a) 显示了第一卷积的滤波器层是由暹罗网络学到的。此外，图6 (b) 显示了第一层子集的左右部分2通道网络2ch的滤波器。值得一提相应的左右部分看起来像是彼此为负，这基本上意味着工作已经学会了在以下情况下计算特征差异：介于两个补丁之间（不过请注意，并非所有第一层2ch的滤波器看起来像这样）。最后，我们在图7中显示一些排名最高的错误和正确匹配项，由2ch深度网络。我们观察到假匹配可能甚至容易被人误解（例如，注意，误报示例中的两个补丁看起来如何相似喜欢）。

在其余的实验中，我们注意到我们使用mod-对Liberty数据集进行培训的els。

培养	测试	2ch-2流	2ch深	2路	暹	暹罗2	伪暹罗语	伪暹罗-1-2	siam-2stream	siam-2stream-l2	[19]
尤斯	ND	2.11	2.52	3.05	5.75	8.38	5.44	8.95	5.29	5.58	6.82
尤斯	解放	7.2	7.4	8.59	13.48	17.25	10.35	18.37	11.51	12.84	14.58
ND	尤斯	4.1	4.38	6.04	13.23	15.89	12.64	15.62	10.44	13.02	10.08
ND	解放	4.85	4.55	6.05	8.77	13.24	12.87	16.58	6.45	8.79	12.42
解放	尤斯	5	4.75	7	14.89	19.91	12.5	17.83	9.02	13.24	11.18
解放	ND	1.9	2.01	3.03	4.33	6.01	3.93	6.58	3.05	4.54	7.22
意思		4.19	4.27	5.63	10.07	13.45	9.62	13.99	7.63	9.67	10.38



表1.几种模型在“本地图像补丁”基准上的性能。模型架构如下: (i) 2ch-2stream  
由两个分支C (95, 5, 1) -ReLU-P (2, 2) -C (96, 3, 1) -ReLU-P (2, 2) -C (192, 3, 1) -ReLU组成-C (192, 3, 1) -ReLU, 一个代表cen-  
tral和一个环绕声部分, 然后是F (768) -ReLU-F (1) (ii) 2ch-deep = C (96, 4, 3) -Stack (96) -P (2, 2) -Stack (192) -F (1) ,  
其中Stack (n) = C (n, 3, 1) -ReLU-C (n, 3, 1) -ReLU-C (n, 3, 1) -ReLU。 (iii) 2ch = C (96, 7, 3) -ReLU-P (2, 2) -C (192, 5, 1) -ReLU-  
P (2, 2) -C (256, 3, 1) -ReLU-F (256) -ReLU-F (1) (iv) 暹罗有两个分支C (96, 7, 3) -ReLU-P (2, 2) -C (192, 5, 1) -ReLU-P (2, 2) -  
C (256, 3, 1) -ReLU和决策层F (512) -ReLU-F (1) (v) siam-l 2减少为siam的单个分支 (vi) 伪siam已解耦



Original text

C(256, 3, 1)-ReLU and decision layer F(512)-ReLU-F(1) (v) siam-l2 reduces to a single branch of siam (vi) pseudo-siam is uncoupled

[Contribute a better translation](#)

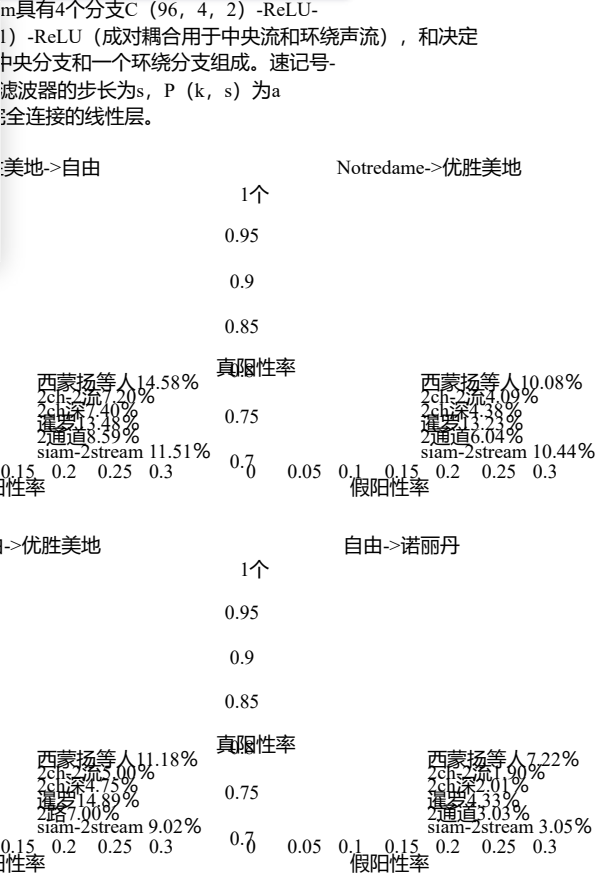


图5.各种模型的ROC曲线（包括最新的描述符[19]）在本地映像补丁程序基准上。号码图例中的是对应的FPR95值

5.2. 广泛的基准立体声评估

为了进行评估,我们选择了Strecha等人的数据集。[\[23\]](#), 其中包含几个带有地面的图像序列  
真相法和激光扫描深度图。我们用了  
“喷泉”和“herzjesu”序列产生6和5个rec-  
固定立体声对。两个序列中的基线

我们选择随着每个图像的匹配而增加  
更加困难。我们的目标是证明光度成本  
用神经网络计算的计算机可与之抗衡  
最先进的手工制作功能所产生的成本  
脚本编写者, 因此我们选择与DAISY [\[26\]](#)。  
  
由于我们的重点不是效率, 因此我们使用了

第7页

用于计算光度成本的优化管道。  
更具体地说, 对于2频道网络, 我们使用了蛮横的  
强制方法, 我们在对应的位置提取补丁  
具有亚像素估计的对极线, 构建批次  
(包含来自左侧图片I 1的补丁和所有补丁  
在右边图像I 2的相应对极线上)  
并计算网络输出, 导致成本:

$$C(p, d) = -\text{onet}(I_1(p), I_2(p+d)) \tag{2}$$

在此,  $I(p)$  表示邻域强度矩阵

2通道架构生成的深度图。结果  
没有全局优化也表明  
深度图比DAISY包含更多的细节。  
对于以下情况, 它们可能会显示出非常稀疏的错误集  
基于暹罗的网络, 但是这些错误非常容易  
在全局优化过程中消除了。  
图8还显示了定量比较, 重点放在  
这种情况适用于基于暹罗语的模型, 因为它们效率更高-  
科学的。该图的第一幅图显示 (对于单个立体声  
对) 偏离地面真相的分布  
跨越所有错误阈值范围 (此处表示为

像素周围的邻域 $\mathcal{N}(p_1, p_2)$ ，是神经元的输出极线上的点之间的距离。

对于暹罗型网络，我们为两个图像中的每个像素一次，然后将它们与决定顶层或 $l_2$ 的距离。在第一种情况下，公式光度成本如下：

$$C(p, d) = -\text{顶部}(D_1(I_1(p)), D_2(I_2(p+d))) \quad (3)$$

其中 $o_{\text{top}}$ 是最高决策层的输出， $D_1, D_2$ 是暹罗或伪暹罗的分支的输出网络，即描述符（对于暹罗网络 $D_1 = D_2$ ）。对于 $l_2$ 匹配，它满足：

$$C(p, d) = D_1(I_1(p)) - D_2(I_2(p+d))^2 \quad (4)$$

值得注意的是，以上所有成本均可计算使用与[28]。这实质上意味着要处理所有完全连接的 $cr$ 为 $1 \times 1$ 卷积，计算暹罗分支网络仅一次，并进一步计算输出这些分支以及网络的最终输出在所有地点使用全程即时通行证（例如，对于2通道架构，计算光度成本仅需输入使用大小为 $s_2 \cdot d$ 的最大全2通道图像对网络进行设置等于输入图像对，其中 $s$ 是第一个步幅网络层和 $d_{\text{max}}$ 是最大差异）。一旦计算，光度成本随后在以下成对的MRF能量中用作一元项

$$E(\{d_p\}) = \sum_p C(p, d_p) + \sum_{(p, q) \in E} (\lambda_1 + \lambda_2 \exp(-\frac{\|D(p) - D(q)\|}{\sigma})) \cdot |d_p - d_q|,$$

使用算法最小化[8]基于FastPD [12]（我们组 $\lambda_1 = 0.01, \lambda_2 = 0.2, \sigma = 7$ 和 $E$ 是4-连接的网格）。

我们在图9中显示了一些定性结果计算深度图（有或没有全局优化”（喷泉）图像集（“herzjesu”ap的结果-梨在补充。缺少空间）。全球MRF优化结果直观地验证了光度学成本用神经网络计算的结果比具有手工制作的功能以及高质量的

场景深度范围的一部分）。此外，其他同一图的图形总结了相应的分布六个立体对对的误差分布行（在这种情况下，我们还分别显示了错误分布-仅考虑未遮挡的像素）。在这些图中，错误阈值设置为3和5像素（请注意，最大视差在500最大基线）。可以看出，所有暹罗车型都在所有错误阈值上，形式都比DAISY更好和所有基线距离（例如，注意相应曲线的曲线）。

5.3。局部描述符性能评估

我们还在Mikolajczyk数据集上测试了网络，CAL描述符评估[16]。数据集包含48相机视点变化，模糊，压缩，照明变化和变焦，逐步变形量不断增加。有已知的理由彼此之间的真相单应性按顺序。

测试技术与[16]。简要地，测试一对图像，将检测器应用于两个图像提取关键点。关注[10]，我们使用MSER检测器。检测器提供的椭圆用来提取输入图像中的补丁。椭圆尺寸被放大3的系数以包含更多上下文。然后，取决于网络的类型，可以是描述符，也可以是暹罗或伪暹罗分支，已提取或全部将补丁对分配给2通道网络以分配分数。

显示了此数据集上的定量比较图10中的几种模型。在这里，我们还测试了CNN网络siam-SPP- $l_2$ ，它是基于SPP的暹罗语体系结构（请注意，siam-SPP与siam相同，但加上两个SPP层-另请参见图4）。我们使用插入的SPP层，其空间尺寸为可以看到 $4 \times 4$ 。的性能，这表明此类Ar-比较图像补丁时的结构。关于休息模式，在图中观察到的结果10 recon-确定先前实验得出的结论。我们只是再次注意到非常好的表现siam-2stream- $l_2$ 的（尽管未接受过

第8页

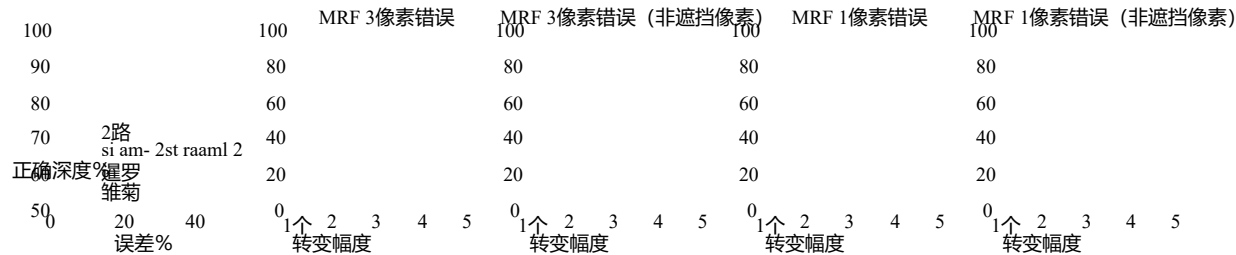


图8.“喷泉”数据集上的宽基线立体声的定量比较。（最左图）与地面的偏差分布真实度，表示为场景深度范围的一部分。（其他图）基线增加的立体声对的误差分布（水平轴），无论是否考虑到被遮挡的像素（在这些图中，错误阈值均设置为等于1和3像素-最大值视差约为500像素）。

图9.宽基线立体声评估。从左至右: DAISY, siam-2stream-l 2, siam, 2ch。第一行-“赢家通吃” depthmap, 第二行-经过MRF优化的深度图。

12个距离)能够明显胜过SIFT和也与图像网络训练的功能的性能相匹配(不过,使用的是较低的512维)。

## 6. 结论

在本文中,我们展示了如何直接学习原始图像像素是补丁的通用相似度函数,以CNN模型的形式进行编码。为了那个最后,我们研究了几种神经网络架构,特别适合此任务,并表明他们表现出非常好的性能,明显优于在几个问题上形成了最新技术标记数据集。

在这些架构中,我们注意到基于2通道的就结果而言,那些显然是优越的。它是,因此,值得研究如何进一步加速未来对这些网络的评估。看待-基于暹罗的架构,2流多分辨率型号非常坚固,可以提供大大提高了性能并验证了比较时多分辨率信息的重要性补丁。同样的结论适用于基于SPP的暹罗语网络,这也不断提高了结果<sup>1</sup>。

实际上, SPP性能可以进一步提高,因为没有倍数在训练SPP模型(例如补丁仅在测试时出现)。

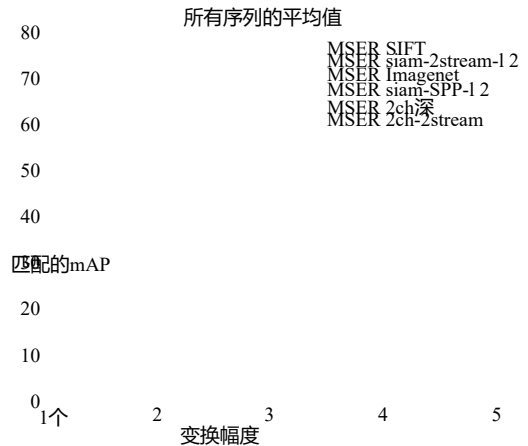


图10.对Mikolajczyk数据集的评估[16]显示所有类型交易的平均平均精度 (mAP) 数据集中的记录(通常, mAP分数用于测量面积在精确调用曲线下)。提供了更详细的图在补充材料中由于空间不足。

最后,我们应该注意,简单地使用较大的火车集合可以潜在地受益并改善总体绩效我们的方法更进一步(作为培训集)在本实验中使用的实际上可以是(以今天的标准来算是相当小的)。

## 第9页

## 参考文献

- [1] X. Boix, M. Gygli, G. Roig and L. Van Gool. 稀疏数量补丁说明的标准化。在CVPR中, 2013年<sup>1, 5</sup>
- [2] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger and R. Shah. 使用“暹罗”时间延迟神经网络进行签名验证网络。在NIPS, 1994. <sup>2</sup>
- [3] M. Brown, G. Hua and S. Winder. 区分学习本地图像描述符。IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010年<sup>2, 4, 5</sup>
- [4] M. Brown, G. Hua and S. Winder. 区分学习本地图像描述符。模式分析与机器情报, IEEE Transactions, 33 (1) : 43-57, 2011年1月。 <sup>4</sup>
- [5] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. 卡坦扎罗和E. Shelhamer. cudnn: 有效的原语进行深度学习。CoRR, abs / 1410.0759, 2014年<sup>4</sup>
- [6] S. Chopra, R. Hadsell and Y. LeCun. 学习相似性判别指标, 适用于人脸验证。在CVPR中, 2005年<sup>2</sup>
- [7] R. Collobert, K. Kavukcuoglu and C. Farabet. 火炬7: A 类似于Matlab的机器学习环境。在BigLearn中
- [18] AS Razavian, H. Azizpour, J. Sullivan and S. Carlsson. CNN具有现成的功能: 惊人的基线获得认可。在IEEE计算机视觉会议上和模式识别, 2014年CVPR研讨会, 哥伦布, 美国俄亥俄州, 2014年6月23日至28日, 第512-519页, 2014年<sup>2</sup>
- [19] K. Simonyan, A. Vedaldi and A. Zisserman. 学习本地使用凸优化的特征描述符。IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014年<sup>2, 5, 6</sup>
- [20] K. Simonyan and A. Zisserman. 非常深的卷积用于大规模图像识别的国家网络。CoRR, ABS / 1409.1556, 2014年<sup>3</sup>
- [21] N. Snavely, SM Seitz and R. Szeliski. 摄影旅游: 探索3d中的照片集。ACM Trans. 图形, 25 (3) : 835-846, 2006年7月<sup>1</sup>
- [22] N. Snavely, SM Seitz and R. Szeliski. 建模互联网照片集的世界。诠释 J. 计算机 视觉, 80 (2) : 189-210, 2008年11月<sup>1</sup>
- [23] C. Strecha, W. von Hansen, LJV Gool, P. Fua and 汤尼森 (U. 在基准相机校准和多视图立体声, 可用于高分辨率图像。在CVPR中。



- NIPS研讨会, 2011年[1, 4](#)
- [8] B. Conejo, N. Komodakis, S. Leprince和J.-P. Avouac. 通过学习进行推理: 加速图形模型优化通过修剪分类器的从粗到细级联来实现。在NIPS中, 2014年[1, 7](#)
- [9] D. Eigen, C. Puhrsch和R. Fergus. 深度图预测使用多尺度深度网络从单个图像中提取。在NIPS, 2014年[1, 2](#)
- [10] P. Fischer, A. Dosovitskiy和T. Brox. 描述符匹配与卷积神经网络: 与SIFT的比较。CORR, ABS / 1405.5769, 2014年[2, 4, 7](#)
- [11] K. He, X. Zhang, S. Ren和J. Sun. 空间金字塔池在深度卷积神经网络中用于视觉识别。在ECCV14, 第III页: 346-361, 2014年[1, 4](#)
- [12] N. Komodakis, G. Tziritas和N. Paragios. 快速, 大约单个和动态MRF的最佳解决方案。在CVPR, 2007年[1, 7](#)
- [13] A. Krizhevsky, I. Sutskever和GE Hinton. 影像网深卷积神经网络进行分类。在F. Pereira, C. Burges, L. Bottou和K. Weinberger, editors, 神经信息处理系统的进步25, 第1097-1105页。柯伦Associates公司, 2012年[1, 2](#)
- [14] Y. LeCun. 反向传播的理论框架。在1988年连接主义模型夏季会议论文集中学校, 1988年第21-28页[1](#)
- [15] DG Lowe. 尺度不变的特异性图像特征关键点。国际计算机视觉杂志, 60: 91-110, 2004年[1, 2](#)
- [16] K. Mikolajczyk和C. Schmid. 绩效评估本地描述符。IEEE模式分析交易 & 机器学习27 (10) : 1615年至1630年, 2005年[2, 7, 8](#)
- [17] E. Nowak和F. Jurie. 学习视觉相似度确保比较从未见过的对象。在CPVR 2007中-IEEE计算机视觉与模式识别会议, 第1-8页, 美国明尼阿波利斯, 2007年6月。IEEE计算机学会。[1个](#)
- [24] C. Szegedy, W. Zarembka和I. Sutskever, J. Bruna, D. Erhan, IJ Goodfellow和R. Fergus. neu-的有趣特性ral网络。CoRR, abs / 1312.6199, 2013年[1, 2](#)
- [25] Y. Taigman, M. Yang, M. Ranzato和L. Wolf. 底面: 缩小人脸验证中与人类水平表现的差距。在计算机视觉与模式识别会议上国家 (CVPR) , 2014年[1, 2](#)
- [26] E. Tola, V. Lepetit和P. Fua. 的快速本地描述符密集匹配。在计算机视觉与程序模式识别, 阿拉斯加, USA, 2008年[2, 6](#)
- [27] T. Trzcinski, M. Christoudias, V. Lepetit和P. Fua. 学习-使用Boosting-Trick来添加图像描述符。在NIPS中, 2012年[2](#)
- [28] J. Zbontar和Y. LeCun. 计算立体声匹配卷积神经网络来降低成本 CoRR, ABS / 1409.4326, 2014年[2, 7](#)